1-2021

An Automated Method to Enrich and Expand Consumer Health Vocabularies Using GloVe Word Embeddings

Mohammed Ibrahim
University of Arkansas, Fayetteville

An Automated Method to Enrich and Expand Consumer Health Vocabularies Using GloVe Word Embeddings

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Engineering with a concentration in Computer Science

by

Mohammed Ibrahim
University of Anbar
Bachelor of Science in Computer Science, 2008
University of Anbar
Master of Science in Computer Science, 2011

July 2021
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____
Susan Gauch, Ph.D.
Dissertation Director

_____
David Andrews, Ph.D.
Committee Member

_____
Brajendra Nath Panda, Ph.D.
Committee Member

_____
Paul Cronan, Ph.D.
Committee Member

Abstract

Clear language makes communication easier between any two parties. However, a layman may have difficulty communicating with a professional due to not understanding the specialized terms common to the domain. In healthcare, it is rare to find a layman knowledgeable in medical jargon, which can lead to poor understanding of their condition and/or treatment. To bridge this gap, several professional vocabularies and ontologies have been created to map laymen medical terms to professional medical terms and vice versa. Many of the presented vocabularies are built manually or semi-automatically requiring large investments of time and human effort and consequently the slow growth of these vocabularies. In this dissertation, we present an automatic method to enrich existing concepts in a medical ontology with additional laymen terms and also to expand the number of concepts in the ontology that do not have associated laymen terms. Our work has the benefit of being applicable to vocabularies in any domain.

Our entirely automatic approach uses machine learning, specifically Global Vectors for Word Embeddings (GloVe), on a corpus collected from a social media healthcare platform to extend and enhance consumer health vocabularies. We improve these vocabularies by incorporating synonyms and hyponyms from the WordNet ontology. By performing iterative feedback using GloVe's candidate terms, we can boost the number of word occurrences in the co-occurrence matrix allowing our approach to work with a smaller training corpus.

Our novel algorithms and GloVe were evaluated using two laymen datasets from the National Library of Medicine (NLM), the Open-Access and Collaborative Consumer Health Vocabulary (OAC CHV) and the MedlinePlus Healthcare Vocabulary. For our first goal, enriching concepts, the results show that GloVe was able to find new laymen terms with an F-score of 48.44%. Our best algorithm enhanced the corpus with synonyms from WordNet,

outperformed GloVe with an F-score relative improvement of 25%. For our second goal, expanding the number of concepts with related laymen's terms, our synonym-enhanced GloVe outperformed GloVe with a relative F-score relative improvement of 63%.

The results of the system were in general promising and can be applied not only to enrich and expand laymen vocabularies for medicine but any ontology for a domain, given an appropriate corpus for the domain. Our approach is applicable to narrow domains that may not have the huge training corpora typically used with word embedding approaches. In essence, by incorporating an external source of linguistic information, WordNet, and expanding the training corpus, we are getting more out of our training corpus. Our system can help building an application for patients where they can read their physician's letters more understandably and clearly. Moreover, the output of this system can be used to improve the results of healthcare search engines, entity recognition systems, and many others.

Acknowledgments

First and foremost, I would like to thank Allah (glory be to Him) for the endless help and the spiritual guidance to all my life steps.

Profound thanks to the one I will never forget her endless help during my Ph.D. journey, Dr. Susan Gauch. Without her guidance and ongoing support, this dissertation would not have been possible.

To my parents, I would say thanks for all that you did in my life. Your support, hard work, and supplication made this achievement possible.

To my wife, Israa, and my kids, Mina, Mustafa, and Lina: I would express my gratitude for your love, care, and support during this journey. I am blessed to have you.

Besides, my gratitude is to my sponsor, the Higher Committee for Education Development in Iraq. Your scholarship made my Ph.D. dream come true.

Furthermore, I would like to thank my committee members, Dr. David Andrews, Dr. Brajendra Nath Panda, and Dr. Paul Cronan, for your constructive comments, guidance, and notes to do my dissertation in the best form. My thanks to our university faculty and staff for their boundless help and commitment to the university and students.

Dedication

To my parents, Salah and Saadiya

To my wife, Israa

Table of Contents

List of Figures

List of Tables

1 Introduction

As the majority of medical terms come from Greek and Latin [1], the medical community has put a lot of effort into translating these terms into English and other common languages. This effort has resulted in clear explanations and definitions for those terms, making the teaching of biomedical concepts in schools and universities much clearer. However, these English medical terms are still hard for laymen to understand (e.g. patients). Terms like *coronary*, *pulmonary* and *oedema*, which are medical jargon written in English, are obvious to biomedical experts, but they are obscure to laymen. Replacing these terms with *heart*, *lung*, and *swelling* would make them easily understood by experts and laymen alike[2]. Mapping medical terminology into clear and easy terms could increase the effectiveness of communication between medical experts and laymen.

With the advancement of medical technology and the emergence of Internet social media, people are more connected than before. In terms of medical technology, there are many efforts to build smart devices that can interact and provide health information. There are already several chatbots designed to interact with people in different medical and healthcare problems. Doctor Apollo[3], Dr. Vdoc[4], and MedBot[5] are all chatbots designed to provide information to patients. Nowadays, many users do not even write what they are looking for. They can just say it, and their smart device can replay with the best-related answers. In terms of social media, people started not only sharing their climate concerns, politics, or social problems, but also their health problems. There are many healthcare social media that provide online consultations for patients. The Pew Research Center reported that in 2011 that about 66% of Internet users looked for advice or opinions regarding their health issues [6]. The rate of using social media by physicians also grew from 41% in 2010 to 90% in 2011 [7]–[9]. In all these cases, doctors, health chatbots

and robots will not be able to interact effectively with laymen unless they have a lexical source or ontology that defines medical jargon.

Recently, steps have been taken to close the gap between the vocabulary that the experts in healthcare are using and what laymen use. In 2017, the Academy of Medical Royal Colleges, which has about 250,000 British doctors, started an initiative in which the doctors asked to write to patients directly using plain English instead of medical jargon [10]. It was reported by [11] that approximately five million doctor letters are sent to patients each month. Using words like *liver* instead of *hepatic*, *brain* instead of *cerebral*, and *children* instead of *pediatric* would make the doctor's letters much easier to laymen than using complicated medical terminology [2].

A concentrated effort, involving experts in different health fields, has created several electronic medical vocabulary resources. Medical resources, commonly known as ontologies, such as Mesh, SNOMED CT, RxNorm, and many others, were built to define, describe, and connect, as much as possible, all the medical jargon. The United States National Library of Medicine (NLM) combined more than two hundred health resources into one thesaurus called the Unified Medical Language System (UMLS). UMLS is a metathesaurus consisting of more than 3,800,000 professional biomedicine concepts. It lists biomedical concepts from different resources, including their part of speech and variant forms. It also defines the relationships between these concepts and arranges the concepts into hierarchies. The UMLS concepts are also classified into two hundred semantic categories, such as Disease and Symptoms, Clinical Drugs, Organisms, and many others. The goal of UMLS is to help computers understand the language of biomedicine [12], [13].

In contrast to the UMLS, Open-Access and Collaborative Consumer Health Vocabulary (OAC CHV), is a collection of medical terms written in plain English. It provides a list of

simple, easy, and clear terms that laymen prefer to use to refer to a professional medical term. The goal of developing these laymen terms is to lessen the gap between laymen and medical experts and to improve the accuracy of health information retrieval [14]–[16]. OAC CHV vocabulary was built manually by a group of experts in the field of biomedicine. The last official update to the this vocabulary was in 2011 in a step to automate the process of extracting laymen terms [16]. NLM has integrated and mapped this vocabulary to UMLS. Out of 3,800,000 concepts on the UMLS, only 56,000 professional medical concepts have been assigned a laymen term(s). Since OAC CHV covers only 1.4% of the available UMLS concepts, there is a lot of work still to be done to enhance it.

Another consumer health vocabulary that the NLM has also integrated into the UMLS is the MedlinePlus health vocabulary. This vocabulary was built as an indexing source for the MedlinePlus search engine [17]. This vocabulary is different from the OAC CHV vocabulary in that the NLM updates this source yearly. The number of laymen terms associated to the UMLS professional medical concepts grew from 2,112 terms in the UMLS version 2018 to 2,140 terms in the UMLS 2020. Even though the OAC CHV vocabulary is bigger than the MedlinePlus vocabulary, the last one has the advantage of growing and updating yearly.

During our investigations of these two consumer health vocabularies, we found that they have many issues. For the OAC CHV vocabulary, we found that out of the 56,000 laymen terms assigned to UMLS professional concepts, 27,000 concept's terms are still jargon and are just repetitions of the professional medical concept. The only difference is that the laymen terms contain either downcased letters, the plural 's', or numbers and punctuations that have been removed from the professional concept. Thus, almost half or 48% of the already assigned laymen terms are still jargon. This shows that the problem is not only adding new terms to non-laymen

concepts but also to enrich the already assigned laymen terms. Similarly, we found that many of the MedlinePlus laymen terms are the same as their associated professional medical concepts. We also found that between the UMLS version 2018 and UMLS version 2020, there were only 28 new MedlinePlus laymen terms associated to their UMLS concepts. The low coverage rate for UMLS concepts is understandable because the identification of, and mapping of, these laymen terms is semi-manual, and this costs time and experts' effort. An automated system to boost the process of finding new laymen terms and mapping them to their UMLS concepts is recommended to eliminate the need for human intervention and avoid wasting time.

1.1 Goals

To address the issues identified above, we propose a system that has the ability to automatically identify new laymen terms and map them to the UMLS. These new laymen terms will either be appended to an already existing list of concept's associated laymen terms, or they will be added to the previously unmapped professional concepts in the UMLS. Our proposed system will tackle two goals:

1.1.1 Goal 1: Enriching Existing Laymen Terms

This goal addresses the problem of identifying new *laymen terms* to add to already existing laymen terms for a given concept. To do that, we build a list of seed terms from the already existing vocabularies, OAC CHV and MedlinePlus vocabulary. For each seed term, we collected instances of its associated laymen terms from one of the commonly used healthcare social media platforms, which is Medhelp.org. Our methods use a combination of Natural Language Processing (NLP) techniques (e.g., data cleaning, tokenization, stemming) to preprocess the collected text into our training corpus. Then, the Global Vectors for word Representations (GloVe) is applied to identify candidate terms from the contexts in which the laymen terms

appear. GloVe is one of word embedding algorithms that takes very large corpora as input to produce a list of word vector representations. From these vectors, a list of seed and candidate term pairs are listed using cosine similarity measurement. This list is ranked, and the highest ranked terms are considered as new laymen terms.

GloVe deals with very large corpora to find the best word representations. Sometimes it is hard to find such large corpora in domain-specific areas, such as healthcare. Due to that, we decided to enhance the GloVe algorithm in four ways. Three approaches leverage the English lexical resource, WordNet. This resource provides many word relations such synonyms, meronyms, antonyms, hypernyms, and hyponyms. We used three of these relations, which are synonyms, hypernyms, and hyponyms to expand our healthcare corpus with new terms. For the synonym approach, a list of synonyms for every seed term occurs in corpus is extracted from the WordNet and added into the context of that seed term. The same process for the hypernym and hyponym approaches is applied where a list of hypernyms or hyponyms is extracted and added to the context of seed term. The other approach uses the idea of iteratively feeding back GloVe's candidate terms again to GloVe co-occurrence matrix. In this approach, a list of top ranked candidate terms extracted from GloVe and iteratively fed back to GloVe to increase the chance of finding new laymen terms. More details about GloVe and its enhancements are reported in Chapter 3. The results of applying these approaches and a comparison of the best methods are reported in Chapter 4.

GloVe and our novel algorithms were evaluated using two laymen datasets from the National Library of Medicine (NLM), Open-Access and Collaborative Consumer Health Vocabulary (OAC CHV) and MedlinePlus Healthcare Vocabulary.

1.1.2 Goal 2: Expanding the Set of Professional Medical Concepts that do not have Associated Laymen Terms

Our second goal tackles the problem of a*dding laymen terms to professional concepts that do not have any existing laymen terms to use as seeds*. There are two issues that this goal needs to find a solution to. First, can we find non-laymen terms from these concepts to use as seeds based on which we can find their associated laymen terms? Second, what is the source of data from which we can find new laymen terms? The answer to the first question is that we will use the UMLS ontology because it has many professional concepts and only 1.4% of these concepts have been mapped to their associate laymen term(s). For the second issue, we will use the MedHelp.org corpus and if it is not a good source of professional medical concepts, we will collect a new corpus from different healthcare resources. More details of this goal are reported in Chapter 3.

To evaluate this goal, we will use the best approach that goal one will report. This goal will use the same ground truth dataset leveraged in goal one. However, in the second goal, the concepts are the input to the proposed approach instead of their associated laymen terms. the concept's laymen terms will be the results that our method should find.

2 Literature Review

2.1 Ontologies

An ontology is a formal description and representation of concepts and their definitions, relations, and classifications in specific or general domains of discourse [18]. It can decrease terminological and conceptual confusion between system software components and facilitates interoperability. Ontologies provide a shared understanding of concepts by defining not only concept synonyms but also their semantic relations (e.g., is-a, part-of, leads to, causes, …etc) [19]. The essential element of any ontology is the concept. A concept in an ontology can have associated terms by which the concept is discussed in text that indicate instances of that concept. An example of a concept and its terms is the concept *Car* that has associated terms such as *jeep, taxi*, and *sedan*. An ontology can be viewed as a source of controlled vocabulary that provides a complete description and interpretation of a concept and its relations in a hierarchical way [20], [21]. An example of interrelated concepts are the concepts *Person*, *Researcher*, and *Manager*. The last two concepts can be considered siblings, both inheriting from the superconcept *Person* [21]. Ontologies provide a simple and abstract representation of what is more than a term, which is a concept, providing its relations, taxonomies, properties, and its instances [22]–[24].

Ontologies simplify the process of text processing and information retrieval by providing mappings from specific terminology to abstract concepts that are useful to applications such as query expansion and question answering [25]. The Semantic Web, which is an Internet technology that aims to make data readable not only by a human but also by computers, uses the ontologies as a main source to understand data and provide information about it [26]. To support the Semantic Web, many ontologies have been created, e.g., BabelNet [27], Disease Ontology [28], Arabic Ontology [29], WordNet [30], Gene Ontology [31], and UMLS [32]. Ontologies

have been used in many domains such as document indexing [33], [34], personalizing user's profiles for information retrieval systems [35]–[40], providing readable data for semantic web applications [41]–[44], and providing interoperability between software systems that exchanging data[45].

2.2 Ontology Creation

The past few years have witnessed a high demand for building ontologies in different domains [46]. According to (Gruber, 1995), any ontology should comply with criteria such as clarity, coherence, and extensibility to be considered as a source of knowledge that can provide shared conceptualization [47]. There are three mechanisms to build ontologies (manual, automatic, and semiautomatic creation) [48]. The next three sections explain these mechanisms.

2.2.1 Manual Ontology Creation

Building or updating ontologies manually can produce precise, coherence, and reliable ontologies. Due you to that, many ontology editors have been proposed such as Apollo [49], OntoStudio [50], Swoop [51], Protégé [52], and OntoGen [53]. Most of these applications provide user-friendly interfaces to all ontology aspects such as classes, relations, functions, and concept attributes [54]. Despite the benefit of manual ontology construction, it has some issues. First, building ontologies manually from scratch, including concept synonyms, attributes, relations, and hierarchies, is immensely time-consuming and requires a lot of human effort. Second, the massive growth of knowledge with the development of electronic resources insists the need for ontologies to be growing too [41], [55], [56].

Having a way to build ontologies automatically or semi-automatically can help reduce the time and labor required to construct ontologies. In the literature, the term *ontology learning* is widely used to refer to the automatic or semiautomatic ontology creation [57]. The following

sections discuss the process of ontology learning and the methods used to enhance and expand ontologies.

2.2.2 Automatic and Semi-automatic Ontology Creation

An automatic or semiautomatic ontology learning is a set of techniques and methods used to build ontologies from scratch, expanding, or enriching already built ontologies with little or no help of a domain expert [58], [59]. Hazman et al. (2011) listed different sources that can be used to learn ontologies, such as structured, semi-structured, and unstructured text [58]. Our research focused on unstructured free text that does not follow any restrictions. Examples of such text include webpages, emails, and social media posts [58], [60]. Typical approaches used in the field of ontology learning are pattern-based methods, Natural Language Processing (NLP) methods, Machine Learning techniques, Text and Data Mining methods, and statistical approaches [58], [59].

Many researches have been proposed to create ontologies from scratch automatically or semi-automatically. Kietz et al. (2000) [71] prototyped an approach to build a company ontology semi-automatically. They started with a general domain ontology called the *GermanNet* ontology, and a dictionary used to extract and classify corporate-related concepts into their taxonomies. After that, a company intranet and public text documents are used to prune unrelated concepts. With the help of human decision, only those concepts that occur on the corporate documents more than the general document were kept in the new ontology.

Farai et al. (2014) introduced an automatic ontology learning system from a domain-specific text. Their system constructed a new ontology using an already built generic ontology. The domain of the text determines the domain of the new ontology. Their system used NLP, Information Extraction (IE), and a lexical database, WordNet [30], to build their domain

ontology. They proved the efficiency of their system using corpus from touristic and legal domains. Once the generic ontology and the corpus input the system, the corpus annotated, and a group of association rules are built that classify each term and populate it into the new ontology. They reported %90 of precision for the legal domain corpus and %76.50 for the touristic domain [61]. There are many other studies and research presented to build ontologies from scratch automatically or semi-automatically such as [62]–[70].

2.2.3 Ontology Modification

The previous section discussed different approaches used to build ontologies automatically or semi-automatically from scratch. Ontologies should continue grow, and new concepts and terms should be added from time to time. Instead of building ontologies from the ground and up, they can be enriched and expanded with new terms and concepts. Many approaches have been proposed to enrich or expand ontologies. The next two sections present some work that has been done in enriching and expanding ontologies in general, and the next section illustrating the research that has been done to build, enrich, and enhance medical ontologies.

2.2.3.1 Ontology Enrichment

Ontology enrichment refers to the process of finding new *terms* to be added to the already built ontology. Agirre et al. (2000) used documents from the internet to enrich the concepts of WordNet ontology. They built their corpus by submitting the concept's senses along with their information, such as their definition, hypernyms, and hyponyms to get the most relevant webpages. They used the AltaVista search engine, currently known as the Yahoo search engine [71]. They used statistical approaches to rank the terms. For every sense document, they computed the word frequency and looked for identical occurring of these words in another sense's documents. These words that repeatedly occur in all sense documents considered to be

synonyms of that sense. The system was able to detect topical signatures for every concept's sense. These signatures can increase the distinction between the concept's senses and help in applications like word disambiguation [72].

Other work in this area has been done by a group at the University of Arkansas. They applied two approaches to enrich ontologies; 1) a lexical expansion approach using WordNet; and 2) a text mining approach. They projected concepts and their instances extracted from already existing ontology to the WordNet and selected the most similar sense using distance metrics. For the text mining enrichment methods, they collected a set of ontology-related documents from the internet using focused-crawling that submits queries to different search engines. Many information extraction methods applied to find and select the most similar terms. The two approaches evaluated using manually created ontology (i.e., an amphibian morphology ontology), finding that text mining approaches provided better performance versus the lexical expansion approaches [73]–[79].

Ali and his team employed multilingual ontologies and documents to enrich not only domain-specific ontologies but also multilingual and multi-domain ontologies. The proposed system consisted of three-level intelligent agents that communicate with each other in a hierarchal way. One agent used to determine the language of the concepts and translates them into different languages. The second one determines the domain of ontologies be enriched, and the last one controls the last agent. The system used the English language as the cross-language source to translate back and forth between other languages. The system tested on the English, German, and Arabic languages using ontologies in the same language domains and multilingual text documents. The proposed system showed a precision of 89% and a recall of 99% [80].

2.2.3.2 Ontology Enhancement

Ontology enhancement is the process of finding new *concepts* to be added to an already existing ontology. Some researchers are using the ontology enhancement, ontology extension, or ontology expansion as an alternative name to the ontology enhancement. Ye et al. (2009) [81] proposed a statistical model to automatically mine concepts from text documents and expand an already built ontology. They applied their system on the ACM CCS ontology [82] and documents collected from the CiteSeer^{x} digital library[83]. Their proposed model showed better performance over other proposed systems in term of efficiency and precision.

Similarly, Tapia-Leon and his team enhanced the BiDO ontology using the NeOn methodology. The BiDo ontology is a scholar-domain ontology that defines the author's bibliographic information such as author's h-index, citations, and journal impact factor [84]. Tapia-Leon used the NeOn scenarios [85] for ontology restructuring and extension, they were able to add new concepts and terms such as author's number of articles, paper citation count, and different other metrics. The proposed ontology validated using OOPS![86] and SPARQL queries submitted using publications from the University of Guayaquil. The validation showed that the new extensions met the standard ontology requirements [87].

The next sections discuss our field of interest. We describe the medical ontologies and why we need them. Furthermore, we illustrate the types of medical ontologies, and the methods and approaches proposed to construct and modify ontologies.

2.2.4 Medical Ontologies

Moving from paperwork to electronic documentation increased the need for a unified medical system that can provide machine-readable data. Moreover, the emphasis on developing an Electronic Health Record (EHR) for patients in the United States encouraged the development of

medical ontologies to ensure interoperability between multiple medical information systems [88], [89]. The medical ontologies consist of the formal or professional vocabularies, the informal, or user-friendly vocabularies, or it can be a combination of both. The next sections discuss these vocabularies and some examples of such vocabularies.

2.2.4.1 Formal Medical Vocabulary

Formal medical vocabularies are all these vocabularies that have been built using formal and professional resources such medical literature, medical codes, genes, drug names and reactions, anatomy, and different diseases [90]. Examples of professional vocabularies are the Disease Ontology [28], Gene Ontology [31], Medical Subject Headings (MeSH) [34], and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [89]. Such ontologies have been applied in many fields such as document indexing for information retrieval systems [34], [91], gene and disease profiling [92], [93], query expansion [94], word sense disambiguation [95], data mining and information extraction [96]–[98], and system software interoperability [99].

2.2.4.2 Informal Medical Vocabulary

Having only professional vocabularies will not help retrieval systems used by laymen. laymen usually use the lay language to express their healthcare concerns. Having an informal health vocabulary, usually known as consumer health vocabulary or user-friendly vocabulary, can loosen this gap and help the human and machine to understand both laymen and professional language. Zeng et al. (2001) reported the poor quality of query retrieval in the professional fields when a layman searched for the term *heart attack*, and reason was because physicians were documenting that concept using the professional concept *myocardial infarction* to refer to that disease [100]. Due to that, many consumer health vocabularies have been proposed, and some of them have been integrated with different professional vocabularies to provide cross-mapped

connection to increase the retrieval rate. Examples of informal medical vocabularies are; the Open-Access and Collaborative Consumer Health Vocabulary (OAC CHV) [16], Italian Consumer-oriented Medical Vocabulary (ICMV) [101], Consumer Health Terminology (CHT) [102], and Chinese consumer health terms (CHT) [103].

Laymen's vocabularies have been applied in many fields such as text simplification [104]–[107], document indexing [108], and identifying Adverse Drug Reactions (ADRs) [109]–[112], . Many of the laymen vocabularies have been mapped or integrated with many professional vocabularies. This indicates that whenever the professional vocabularies used in an application, the consumer vocabularies can be used too. After discussing the medical ontologies and their types and uses, we explain the methods proposed to create, enhance, and enrich such ontologies.

2.2.5 Medical Ontology Creation

Most medical ontologies have been built from the ground and up manually by a group of human experts in the domain of that ontology. The Disease Ontology (DO), for example, is a complete knowledge base of human diseases. It was the result of the tremendous efforts of many domain experts at Northwestern University. This ontology includes disease definitions, relations, and attributes, and not only that, the DO has integrated with many medical ontologies such as SNOMED CT, MeSH, and GO to provide cross-mapped disease information [28]. The 2019 release of this ontology has roughly 9,423 disease concepts, and only 68% of these concepts have been defined [113].

Another manually built medical ontology is the Gene Ontology (GO), a project proposed in 2008 to help annotate genes, genes products, and their sequences. It was an extensive effort of experts at the Gene Ontology Consortium. This ontology includes humans and organism genes

information [31], [114]. The 2016 version of this ontology has approximately 43,585 terms and about 93,265 relationships in different aspects, such as molecular functions, cellular components, and biological processes.

There are many other medical ontologies that have been built manually from scratch such as the Royal Society of Chemistry's name reaction Ontology (RXNO) [115], DrugBank [116], and Human Phenotype Ontology (HPO) [117]. Many other medical ontology examples can be found in [118]. The MedlinePlus vocabulary was constructed to be the source of index terms for the MedlinePlus search engine [17]. The NLM updates this resource yearly. In the UMLS version 2018, there were 2112 professional concepts mapped to their laymen terms from the MedlinePlus topics. Due to the extensive human effort required, there were only 28 new concepts mapped to their associated laymen terms between UMLS version 2018 and UMLS version 2020.

We have already discussed the disadvantages of building ontology from scratch and that an automatic or semi-automatic method is required to solve such problems. Due to the sensitivity of patient information needed for text mining approaches, there have been few projects that attempt to create biomedical ontologies from scratch *automatically*. However, there have been several approaches proposed to construct ontologies semi-automatically. The next section explains the semi-automatic approaches proposed to create medical ontologies.

2.2.6 Semi-automatic Medical Ontology Creation

Semi-automatic ontology creation drove the attention of many researchers due to its ability to build ontologies with less time and human efforts. Not only that, but it can also process a gigantic size of data and present it to the domain expert. Huang et al. (2014) applied knowledge-driven approaches to build miR ontology semi-automatically from scratch using already existing

biomedicine ontologies and miR databases. They started with a seed ontology created manually and extended the seed concepts by searching for them in the other biomedical databases and ontologies. Using artificial neural networks along with some clustering algorithms, they were able to rank concept pairs. A group of domain-experts judged whether the new concepts can be added to the new ontology or not [119].

Similarly, Zeng and his team applied semi-automatic approaches to create the consumer-friendly display (CFD) names. Their approach used a corpus collected from queries submitted to a MedlinePlus website [120], where only frequent expressions are kept for the study. The frequent expressions reviewed by six experts to find their matched UMLS concepts. The review ended with creating the CHD with 400 concepts along with their consumer health terms [121], [122].

Recently, He and his team initiated the coronavirus ontology with the purpose of providing machine-readable terms related to the coronavirus pandemic that occurs in 2020. This ontology includes all related coronavirus topics such as diagnosis, treatment, transmission, and prevention areas [123]. There are several other research has been done to build medical ontologies from scratch semi-automatically that has been discussed in [124]–[128].

2.2.7 Medical Ontology Modification

Medical ontologies need to be updated from time to time due to the growing in the medical knowledge that coming from different sources such medical literature and healthcare social media. Such ontologies need to be enhanced and enriched with new concepts and terms. The following sections discuss the literature that has been introduced to enhance or enrich medical ontologies.

2.2.7.1 Medical Ontology Enhancement

Several research projects focused on enhancing medical ontologies. Pesquita and Couto were able to predict the areas in which the Gene Ontology is more likely to be enhanced. They applied a supervised learning approach to different versions of the Gene Ontology (GO). Their approach can help future research focus on specific areas of the GO ontology and reduce the efforts to develop and update such ontology [129].

The Open-Access and Collaboration Consumer Health Vocabulary (OAC CHV), which is our research of interest, has been enhanced many times by a group at University of Utah. In 2006, the group enhanced the this vocabulary with 1000 UMLS concepts along with their laymen terms using the same semi-automatic methods they applied in their 2005 work[121], [122]. The group continued working on enhancing the OAC CHV, and in 2007 they were able to identify 753 new terms and mapping them to their UMLS concept. They applied the C-value technique [130] and logistic regression methods to extract candidate terms and found that the logistic regression methods were more effective in identifying laymen terms than the C-value technique [131].

In 2011, the group presented a computer-assisted system that processes a text collected from a healthcare social network and displays the candidate's terms to the experts. The system by itself was able to identify 30% of the validation list with human introversion. In this work, roughly 651 terms were defines and mapped to their right UMLS concept [16]. This work was the last official work that the National Library of Medicine (NLM) rely on to enhance OAC CHV vocabulary.

Our enhancement approaches differ from those discussed above in that we developed techniques that are entirely automated without any human intervention. In addition to that, we

used advanced NLP such as word embedding. Finally, we built a complete laymen database that is ready to help finding new laymen terms and evaluate system performance.

2.2.7.2 Medical Ontology Enrichment

The medical ontologies should not only be enhanced but also enriched with new terms. Enrichment usually is easier than enhancement and that is because in the enrichment process the concepts are already defined. In addition to that, these concepts might already have some terms that can be used as seed terms to find their similar pairs. Zheng and Wang (2008) proposed a tool called the Gene Ontology Enrichment Analysis Software (GOEST). It is a web-based tool that uses a list of genes from the GO and enriches them using statistical methods. The tool used data downloaded from different sources such as gene annotated text from gene companies (Agilent, Affymetrix, and Illumina), and definitions and hierarchies of genes downloaded from the Gene Ontology. The GOEST showed more accurate results over many proposed tools and helped determine gene hidden information [132]. Shanavas et al. [133] presented a method to enrich the UMLS concepts with related documents from a pool of professional healthcare documents. Their aim was to provide retrieval systems with more information about medical concepts.

Recently, He and his team at Florida State University enriched the OAC CHV similarity-based technique. They collected posts form a healthcare social media for different healthcare communities and processed these posts using the OpenNLP tool [134]. The OpenNLP processed the texts into n-grams and produced a list of candidates and seed terms. Seeds terms are those terms that have been found in the CHV database. After that, a feature vector is created for every term using ten features such as TF, DF, C-Value, POS, context around the term, and some others. They proposed a method called *simiTerm* that uses the k-means algorithm [135] to find term

pairs. For their evaluation, they sub-sampled some of the seed terms, and their system with it is the best setting showed and F-score around 52% [15].

The drawbacks of this research are that they used the k-means clustering, which already has some disadvantages, such as the initial clusters could impact results, and a user has to specify the number of clusters. Also, the feature set selected empirically, and that could show bias. Finally, their work did not end up adding new laymen terms to the OAC CHV.

Recently, Gu et al. applied three methods to add synonyms to the Chinese professional concepts, i.e., Word2Vec [136], GloVe [137]. and FastText [138] using data collected from two Chinese healthcare communities. In total, they collected 2,180 Gb of text data. The text is filtered to have only those Chinese professional concepts downloaded from the ICD-10[139]. Afterward, they manually collected 224 pairs of professional concepts and their consumer terms along with their context as seed terms. Then, they applied algorithms to the corpus and measured the most similar vectors to the seed terms. They found that the consumer term is, on average, in the 8th place of the most similar term list to the professional concept [140]. Although promising work, the authors did not provide any objective measures to assess the performance of their system. In addition, their approach is semi-automated and requires human effort, which may account for the small size of the dataset used.

Unlike previous approaches discussed here, our proposed system is completely automated. Furthermore, we are not only applying recent techniques but also improving them to increase the chance to find new laymen terms. We also explore a novel approach to selecting the seed terms.

2.3 Natural Language Processing Techniques for Vocabulary Enrichment

2.3.1 Text Mining

The massive growth of electronic text documents and the amount of knowledge and the information that can be found in such documents led to the emerging of the text mining field. It is a broad field that includes different areas such as information extraction, document clustering and classification, ontology construction, data mining, information retrieval, and many other areas [141]–[143]. Rai listed the steps to do any text mining task, which starts with collecting the text, cleaning and remove unnecessary text, applying text mining tools, turning the text into understandable data, and finally analyzing the data and store precious information [141].

Text mining algorithms can be divided into; (1) traditional NLP tools such tokenization, Part of Speech (POS), and named entity recognition approaches. (2) statistical approaches such as Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), word vector representations. (3) machine learning approaches such supervised and unsupervised learning approaches (e.g., Support Vector Machine (SVM), Bayesian and Neural network classifiers, K-means and Expectation Maximization Clustering algorithms) [142]. All these tools and algorithms can be applied individually or together in one application. Text mining can be applied to structured text such as relational databases, unstructured text such as emails and webpages, and on semi-structured text such as delimited text formats.

In our work, we apply a mix of different text mining approaches. The main algorithm in our work is one of the vector representation algorithms called Global Vectors for Word Representations (GloVe) [137]. We apply this algorithm to enrich and enhance CHV vocabulary using unstructured text collected from a healthcare social media.

The next section discusses the word vectors in general and explains the GloVe algorithm. We also discuss the auxiliary resources we used to enhance our GloVe, such as WordNet.

2.3.2 Word Vectors

Word vector representations are a set of models and methods that represent words and their related information arithmetically and turn them into real numbers. These vectors act like features that can be used in different text mining applications. There are many approaches applied to measure the relatedness between word vectors [137]. Typically, distance and angle methods (e.g., Euclidean distance [144] and Cosine distance) are the commonly known methods that can measure that. Recently, Miolov et al. (2013) presented another vector evaluation method that uses word analogies. In their method, they basically applied arithmetic operations between word vectors, such as the vectors of the *queen* - *woman* + *man* = *king* vector [145].

In general, there are two main vector-learning models. The first models incorporate global matrix factorization whereas the second models focus on local context windows. The global matrix factorization models generally begin by building a corpus-wide co-occurrence matrix and then they apply some dimensionality reduction methods. An early example of this type of model is Latent Semantic Analysis (LSA) [146] and Latent Dirichlet Allocation (LDA) [147]. The context-window models are based on the idea that a word can be defined by its surroundings. An example of such models is the skip-gram model [145] proposed by Mikolov in 2013 and the model proposed by Gauch et al. in [148]. Word2Vec [136], FastText [138], and GloVe [137] are all examples of vector learning methods that have been shown to be superior to traditional NLP methods in different text mining applications [140], [149]. Some of these

techniques have been applied in the medical field to build medical ontologies, such as [150]–[154].

In this research, we apply the Global vectors for word representations GloVe algorithm to find new laymen terms and map them to their related UMLS concepts. GloVe is one of the unsupervised learning algorithms that build vectors to statistically describes words in a corpus. It outperformed many vector learning techniques in the task of finding word similarity. It works on the concept that a word can be defined from its surrounding words. GloVe combines the advantages of two vector learning techniques: global matrix factorization methods and local context window methods. It starts with a global word to word co-occurrence matrix. From this matrix, GloVe runs its model to find the best vector representation for every learned word. GloVe uses a global log bilinear regression model to learn word vectors. The model avoids the sparsity of the global co-occurrence matrix by running only on nonzero entries [137]. This algorithm has many applications in different fields such as text similarity [155], node representations [156], emotion detection [157] and many others.

GloVe is generally used with very large corpora, e.g., a 2010 Wikipedia dump (1 billion tokens), a 2014 Wikipedia dump (1.6 billion tokens), Gigaword 5 (4.3 billion tokens) [137]. In comparison, our corpus is specialized and much smaller, approximately 1365,000 tokens. Thus, we have some challenges that require us to develop novel approaches such as incorporating an auxiliary source of vocabulary, e.g., WordNet, or an iterative feedback of top-ranked terms retrieved from GloVe, or a combination of all the above.

WordNet [13] is an online machine-readable English ontology proposed by Professor George A. at the Princeton University. The most recent version has about 118,000 synsets (synonyms) of different word categories such as noun, verb, adjective, and adverb. Wordnet

provides for every synset a short definition and sometimes an example sentence. It also includes a network of relations between its synsets. The synonyms, antonyms, hyponyms, hypernyms, meronymy's and some others are all semantic relations that WordNet provides [30]. WordNet has been used in many fields to help enrich ontologies in different domains such as [158]–[160].

3 Research Plan

The main goal of this research is to automatically identify new laymen terms to either enrich an existing list of laymen terms or add them to a previously unmapped professional medical concept. Although numerous research efforts have been conducted to tackle the problem of identifying new laymen terms automatically, they either did not produce accurate results, or they required expert assistance to add the new terms manually. Many of the proposed systems focused on only two to three broad healthcare community issues (e.g., Diabetes and/or Cancer) and they produced a system that is fairly domain specific and cannot be applied on diverse healthcare topics. In this research, we applied advanced machine learning and NLP tools to extract new laymen terms. Additionally, we evaluated our proposed system on a variety of healthcare topics. Although we are focusing on the medical domain, the techniques that we develop should be applicable to other domains that which to employ word embedding but have limited training data. Figure 3.1 shows the general system architecture of our system.

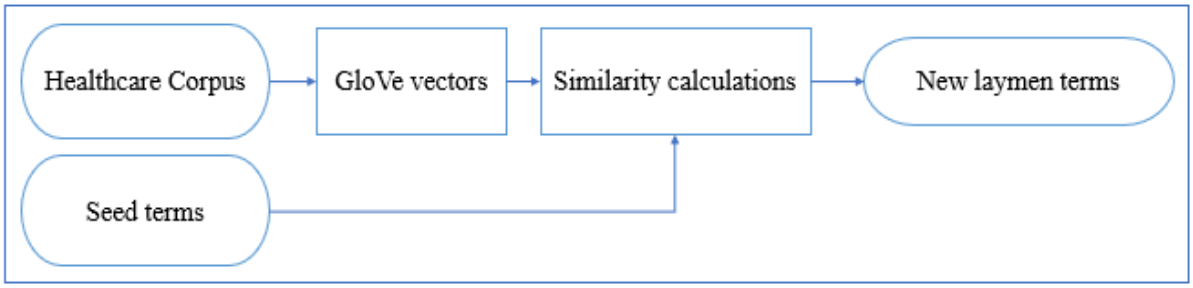

Figure 3. 1 General system architecture

The first section of this chapter discusses the source of healthcare social media that has been utilized to build our corpus. The second section discusses the source of seed terms used to find similar laymen terms. The third section presents in detail our first research goal, i.e., how to enrich the set of laymen terms for previously mapped professional medical concepts. This section explains the word embedding algorithm used to build the best word vectors. It also

includes a discussion of similarity measurement used to rank the candidate terms, and the list of new laymen terms. The final section explains our second goal, i.e., expanding the coverage laymen terms by adding new term(s) to unmapped professional medical concepts.

3.1 Healthcare Social Media Corpus

Medhelp.org is a healthcare social media site in which people post information about their health issues. These posts are presented in a question/answer format wherein people share their experiences, knowledge, and opinions within different health communities. Instead of writing a short query on the Internet that may not retrieve what a user is looking for, whole sentences and paragraphs can be posted on such media [161]. People discussing issues with healthcare on social media tend to use lay language rather than formal medical terminology. For example, they use sentences such as “I can’t fall asleep all night” to refer to the medical term “insomnia” and “head spinning a little” to refer to “dizziness” [162]. Such healthcare social media can be an excellent source from which to extract new laymen terms. We used the *Medhelp.org* [163] as our primary source of lay language to find new laymen terms.

3.1.1 Building our Corpus

The Medhelp.org has hundreds of healthcare communities. To select the communities to include in our corpus, we did an informal experiment to find the occurrences of laymen terms on *Medhelp.org* community. We used the existing laymen terms in the OAC CHV because this vocabulary has a good coverage of laymen terms on the UMLS. We found the highest density of these laymen terms occur in communities such as Pregnancy, Women’s Health, Neurology, Addiction, Hepatitis-C, Heart Disease, Gastroenterology, Dermatology, and Sexually Transmitted Diseases and Infections (STDs / STIs) communities. We downloaded all the questions on these communities to April 20, 2019. The dataset size is roughly 1.3 Gb and

contains approximately 135,000,000 tokens. More details about this corpus and some other statistics related to every downloaded community can be found in Section 4.1.

3.2 Vocabulary Resource

In order to find new laymen terms, our approaches need a list of seed terms for the concepts that indicate contexts in which the new laymen terms might be located in the corpus. Since laymen are most likely to use non-medical terms in the corpus, these seed terms are more likely to be found they themselves are also informal medical terms. The UMLS already contains some layman's medical vocabularies, which are the Open-Access and Collaborative Consumer Health Vocabulary (OAC CHV) and MedlinePlus health vocabulary. These laymen vocabularies list the laymen-friendly medical description for the professional medical concepts. They try to simplify the language between laymen and professionals in the field of biomedicine. Table 3.1 shows some examples of professional medical concepts from UMLS along with their mappings to laymen terms. We used these two vocabularies to build the seed terms that we need to find new laymen terms for. The next section explains the process of getting a list of seed terms from these two laymen vocabularies.

Table 3. 1 Examples of professional medical concepts and their associated laymen terms

No.	Prof. medical concept	Associated laymen terms
1.	Diabetes Mellitus	Diabetes; disorder diabetes mellitus;
2.	Cerebrovascular accident	Apoplexy; cerebral stroke; CVA; stroke; strokes
3.	pneumonia	lung inflammation; inflammation lungs; pneum
4.	Tuberculosis	infection tuberculosis; TB; TBC;
5.	Sleeplessness	insomnia
6.	HIV	AIDS virus; human immunodeficiency virus; htlv iii

3.2.1 Creating a List of Seed Terms

Roughly 56,000 professional medical concepts within UMLS do have existing laymen terms from the OAC CHV vocabulary. Additionally, the MedlinePlus vocabulary provides more than 2,000 UMLS professional medical concepts along with their associated laymen terms. These two vocabularies are good resources of laymen terms to build our seed term list from. We used the UMLS version 2018 to build our seed term list. During our investigating of these two laymen vocabularies, we found some issues. We found that many of the listed laymen terms are the same as the others except adding a plural s, capitalizing letters, changing in term's word order, and some other issues. Table 3.2 shows an example of twelve laymen terms listed from the OAC CHV vocabulary for the *headache* concept.

We can see from Table 3.2 that many of the listed laymen terms are exactly the same as the others. The terms in position 2 and 3 are the same except the plural s. The same for the laymen terms in 6 and 7, and the more trivial ones are these terms in positions 8,9,10,11, and 12. They all refer to the basic form of the term *head pain*.

Table 3. 2 Professional medical concept headache and its associated laymen terms

No.	Laymen terms	No.	Laymen terms
1.	cranial pain	**7.**	cephalgias
2.	headache	**8.**	head pain
3.	headaches	**9.**	head pained
4.	head ache	**10.**	Head Pain
5.	cephalalgia	**11**	head pains
6.	cephalgia	**12**	pain in head

For our research, we need a list of seed terms that has unique and non-trivial laymen terms. To get such terms, we cleaned laymen terms from the stopwords, punctuations, and digits. Then, we lowercased and stemmed all words using the Porter stemmer [164]. We used the stemmer to turn the term to its basic form. Multiple terms with different word order but same

words considered as one. For example, if there are terms like *head pain*, *pain in head*, and *head pains*, the final term will be *head pain*. All duplicates after stemming are removed and only unique terms kept. So, in the case of our example in Table 3.2 for the medical concept *headache*, the final list of its associated laymen terms dropped from 12 laymen terms to 6 laymen terms, which are *{cranial pain, headache, head ache, cephalalgia, cephalgia, head pain}*.

We did this process for the two laymen vocabularies. For the OAC CHV vocabulary, out of the 56,000 professional concepts on the UMLS, only 43,475 professional concepts have non-repeated and non-trivial laymen terms. For the MedlinePlus vocabulary, out of 2,112 professional concepts, only 1,615 concepts have non-trivial laymen terms. A subsample from each list is used to evaluate our first and second goal. Figure 3.2 shows the process of extracting and getting our seed list from the UMLS database for the OAC CHV vocabulary.

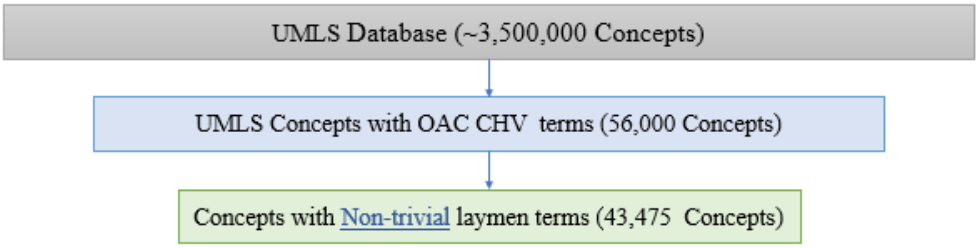

Figure 3. 2 The process of getting non-trivial laymen terms

Table 3.3 shows a sample of some professional medical concepts and their non-trivial laymen terms. These non-trivial terms are used as seed terms to find new laymen terms and enrich their concepts. The CUI in Table 3.3 represents the UMLS concept unique identifier.

Table 3. 3 Professional medical concepts with their associate non-trivial laymen terms.

CUI	Prof. Concept	Non-trivial laymen terms				
C0035334	retinitis pigmentosa	pigmentary	retinopathy	cone	rod	dystrophy
C0035127	repetitive strain injury	cumuli	trauma	motion	overuse	syndrome
C0015252	removal - procedure	excise	extirped	ectomi	surgical	resect
C0034194	pyloric stenosis	stenos	gastric	outlet	obstruct	outflow

3.3 Goal 1: Enriching Existing Laymen Terms

This goal tackles the problem of identifying new laymen terms to be added to professional medical concepts that already have some associated laymen terms. Essentially, we are trying to find synonyms for the seed terms. Our approach is based on using Global Vectors for Word Representation (GloVe) in different ways to identify the new synonyms. The next sections discuss GloVe and its enhancement methods.

3.3.1 Method 1: GloVe (Baseline)

In our first method, we implemented the Global Vectors for Word Representation (GloVe) to get new laymen terms and associate them with their concepts. Figure 3.3 shows the steps of this method. **Step 1** and **Step 3** are already discussed in sections 3.1.1 and 3.2.1 respectively. GloVe is **Step 2** on Figure 3.3. In **Step 2**, GloVe starts collecting word contexts using its global word to word co-occurrence matrix, which is we denoted by *X*. This matrix is a very large and very sparse matrix. Each entry in this matrix has a count for how many times a word *i* occurs in the context of word *j*. Building *X* requires a onetime pass over the whole corpus. This matrix is the primary source of data that GloVe model uses to build word vectors.

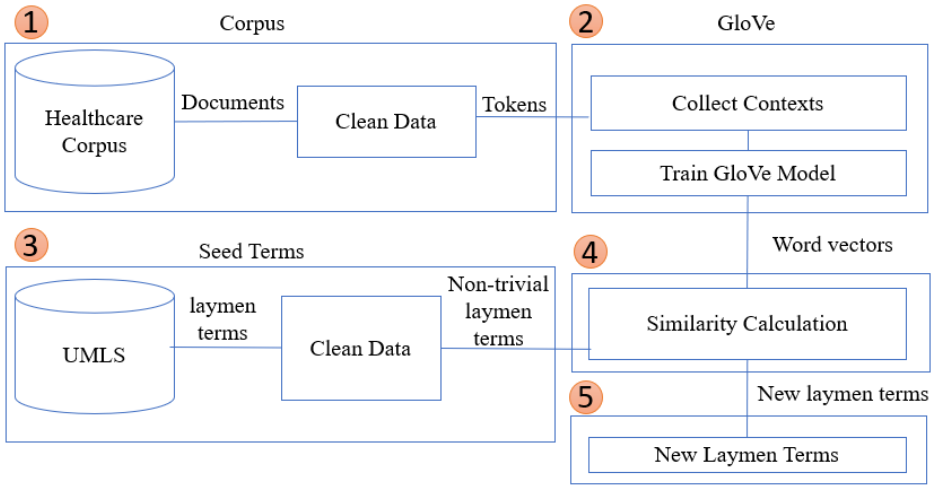

Figure 3. 3 Method one system architecture

Given a word to process, i.e., the *pivot* word, GloVe counts co-occurrences of words around the pivot word within a window of a given size. As the windows shift over the corpus, the pivot words and contexts around them continually shift until *X* is completed. After *X* is built, GloVe builds word vectors for each word that summarize the contexts in which that word was found. These word vectors using the following least squares regression model:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w}_j + b_i + b_j' - \log X_{ij})^2, \qquad (3.1)$$

where:

- *J* is the objective function that tries to find a word vector with the minimum difference between its elements (words).
- *V* is the vocabulary size.
- The function $f(x_{ij})$ plays a primary role in dimensionality reduction. It was reported in [137] that 75%-95% of *X* entries are zeros. It is inefficient to go over such entries and do their computations. The function $f(x_{ij})$ solves this problem by cutting off such computations and returns zero when X_{ij} is equal or close to zero. This function also takes care of the occurrence of rare words and frequent words. It returns 1 when frequent words are overweighed, and a value distributed between 0 and 1 when X_{ij} is less than an *Xmax*. The flowing function describes how this function works:

$$f(x) = \begin{cases} (x/x_{max})^{\alpha}, & if\ x < x_{max} \\ 1, & otherwise \end{cases}, \qquad (3.2)$$

  where *Xmax* is a manually set parameter that acts as a threshold represent a maximum frequency. The α controls the distribution of words that occurs normally on the corpus.
- The $w_i^T \widetilde{w}_j$ is the dot product between vectors of the word *i* and word *j*. The source of these two vectors is the co-occurrence matrix *X*. The b_i and *bj* are scalars that credit word

i and word *j*. The logarithm in the *log* X_{ij} is used to normalize the co-occurrence of *word i* with word *j*, which is expected to be large when the corpus size is large.

GloVe has several other parameters, commonly known as hyperparameters, that can affect its accuracy. The window size and vector dimensionality are parameters that can play a primary role in GloVe results. Table 3.4 shows the best settings that reported in [137].

Table 3. 4 Basic setting for the GloVe hyperparameters

No.	Hyperparameter	Value
1	Window Size	10
2	Vector dimension	300
3	Xmax	100
4	α	3/4

After setting all the required parameters, the GloVe model starts training. The result of that training is a list of vectors with their best representations. After preparing all corpus vectors, **Step 3** and **Step 4** start. In **Step 3**, we use the non-trivial laymen terms as the seed term and submit them to the GloVe model to find their most similar words. In **Step 4**, we implemented the cosine similarity measurement to find the most similar words to every seed term. Cosine similarity measures the angle divergence between two vectors. It produces a score between 0 and 1. The higher the score between two vectors the more identical they are [165]. Equation (3.3) show how we computed the similarity score.

$$\cos\_\text{sim}(v_1, v_2) = \frac{\overrightarrow{v_1}.\overrightarrow{v_2}}{|\overrightarrow{v_1}||\overrightarrow{v_2}|}, \qquad (3.3)$$

where $\overrightarrow{v_1}$ is a vector of a term on the seed term list, and $\overrightarrow{v_2}$ is a vector of a corpus term that GloVe model built. The denominator of Equation 3 is the product of the two vectors lengths. After calculating the similarity score between vectors, a list of the most similar words is listed. Only new laymen terms are added to the professional concepts.

3.3.2 GloVe with WordNet Synsets

Word embedding algorithms usually use a very large corpus to build its word representation, e.g., 6B words of Google News corpus are used to train the word2vec vectors [136], [166]. In the case of a narrow domain such as healthcare, it is hard to find or build an immense corpus, increasing the sparsity of the co-occurrence matrix and impacting the accuracy of the resulting word vectors. Thus, one of our goals is to investigate the ability of an external ontology to increase the accuracy of word embeddings for smaller corpora. In particular, we present methods to exploit a standard English ontology, WordNet, to enhance GloVe's accuracy on a healthcare domain corpus. WordNet provides a network of relations between its relational synsets such as, synonyms, antonyms, hyponyms, hypernyms, meronymy's and some other relations.

In our research, we investigate using the synonym, hyponym, and hypernym relations to augment our corpus prior to running GloVe. We only expand the seed terms in the training corpus with their relational synsets. For each seed term, we located the relational synset of interest, e.g., hyponyms we sort them by similarity to the seed term using the Resnik [167] similarity measurement (See Equation (3.4)).

$$Sim_{resnik}(A, B) = -\log P(LCS(A, B)). \qquad (3.4)$$

where *A* is the seed term and *B* is the relational synset produced from the WordNet ontology. The *LCS* is the Least Common Subsumer (LCS) between two terms *A* and *B*. After having the Resnik score between all the relational synsets and the seed term, we rank the list of synsets according to their Resnik score and then split them evenly into two subsets of roughly equal total similarity using a round-robin algorithm. We then expand the corpus by adding the first subset of relational synset words to the corpus prior to each seed term occurrence and the second subset after each

seed term occurrence. Figure 3.4 shows the methodology of our system with the WordNet ontology. We have highlighted the wordnet method the orange color. All the other steps are the as the ones we explained in the previously discussed GloVe's section.

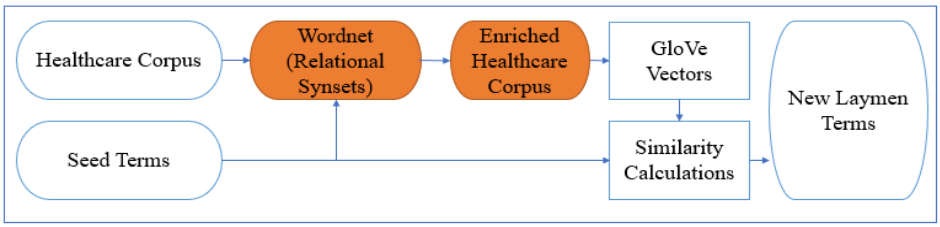

Figure 3. 4 Methodology of improved GloVe with WordNet corpus enrichment

Expressing the WordNet method, let $S = \{s_1, s_2, s_3, \dots, s_n\}$ be a set of *n* seed terms. Let $T = "w_1\ w_2\ w_3\ \dots . w_k\ "$ be a text of words in the training corpus. Let $X = \{x_1, x_2, x, \dots, x_z\}$ be a set of relational synset terms for the seed term s_i, where $i=0,1,2,\dots,n$ These relational synsets are sorted according to their degree of similarity to s_i using the Resnik similarity measurement. *X* is divided into two sets X_1 and X_2 and each set goes to one side of s_i. Now, let $s_i = w_{j+2}$ in T, where $j=0,1,2,3,\dots,k$. Then, the new text $\hat{T}$ after adding the relational synsets will look like this:

$$\hat{T} = "\ w_j\ w_{j+1}\ X_1 w_{j+2} X_2 \dots . w_{j+k}\ "$$

Further, consider the effect of T-hat on the GloVe co-occurrence vectors. Assume that s*i* has the vector $\overrightarrow{V_{s_i}}$. After enriching the training corpus with the relational synsets, the new vector $\widetilde{\overrightarrow{V_{s_i}}}$ will equal:

$$\widetilde{\overrightarrow{V_{s_i}}} = \overrightarrow{V_{s_i}} + \vec{X} \qquad (3.5)$$

The co-occurrence weights of relational synsets that are already in the corpus will be increased incrementally in the vector, while those that are new to the corpus will expand the vector and their co-occurrence weight will be calculated according to the co-occurrence with the seed term. The following sections outline the WordNet approach above with the three types of relational synsets we used: synonyms, hyponyms, and hypernyms.

3.3.2.1 Method 2: GloVe WordNet Synonyms (GloVeSyno)

Synonyms are any words that share the same meaning. For example, the words *auto*, *machine*, and *automobile* are all synonyms of the word *car*. Having synonyms around a seed term adds more information about that seed term and help building more accurate seed term vectors. When a seed term found in the training corpus, WordNet provides a list of its synonyms. These synonyms are sorted according to their degree of similarity to the seed term. After that, the synonyms are divided into two lists and each list go to one side of the seed term. Here is an example that demonstrate this process. Let $T =$ "$I\ had\ a\ headache$" be a text in the training corpus. *T* has the seed term *s* = *headache*. The WordNet synonyms of this seed term are {*concern*, *worry*, *vexation*, *cephalalgia*}. Sorting this set according to their degree of similarity results the following set: {*worry*, *cephalalgia*, *concern*, *vexation*}. This set is divided in to two sets {*worry*, *cephalalgia*} and {*concern*, *vexation*} and added to the left and right of the *s* in *T*. So, the $\hat{T}$ equals:

$$\hat{T} = "I\ had\ a\ worry\ cephalalgia\ headache\ concern\ vexation"$$

Assume that the vector of the seed term *s*, $\overrightarrow{V_s}$, before enriching the training corpus, the vector looks like this:

	dizzy	pain	I	had	a	for	worry	please	sleep
$\overrightarrow{V_s}$	5	0	5	10	1	0	15	0	50

The $\overrightarrow{\tilde{V}_s}$ for the seed term after enriching the training corpus with the WordNet synonyms will be expanded to have the new words and updated the occurrence of the already in corpus words. Here is how the $\overrightarrow{\tilde{V}_s}$ looks like:

	cephalalgia	dizzy	pain	I	had	a	for	concern	worry	please	vexation	sleep
$\overrightarrow{\tilde{V}_s}$	1	5	0	5	10	1	0	1	16	0	1	50

We can see from the $\vec{\tilde{V}_s}$ that the words that are new to the corpus vocabulary expanded the vector and their weights are calculated according to their co-occurrence with the seed term, while the words that are already in the vector, such as *worry*, their weights increased incrementally.

3.3.2.2 Method 3: Glove WordNet Hyponyms (GloVeHypo)

Hyponyms are those words with more specific meaning, e.g.,*Jeep* is a hyponym of *car*. The idea here is to find more specific names of a seed term and add them to the context of that seed term. to explain this method, we use the same example we used in the previous section. The hyponyms of the seed term *headache* that the WordNet provides are {*dead_weight, burden, fardel, imposition, bugaboo, pill, business*}. Sorting these hypos according to their degree of similarity to the seed term results the set {*dead_weight, burden, fardel, bugaboo, imposition, business, pill*}. This list is divided into two sets and each set go to one of the seed term's sides. The rest process is the same as the GloVeSyno method.

3.3.2.3 Method 4: GloVe WordNet Hypernyms (GloVeHyper)

Hypernyms are the antonyms of hyponyms. Hypernyms are those words with more general meaning, e.g., *car* is a hypernym of *Jeep*. The idea here is to surround a seed term with more general information that represents its ontology. Having this information leads to more descriptive vector that represent that seed term. An example of a seed term hypernyms is the hypernyms of the seed term *headache*, which are {*entity, stimulation, negative_stimulus, information, cognition, psychological_feature, abstraction*}. We can see that these hypernyms are broader than the seed term *headache*. We use the same steps for this relational synset as in the GloVeSyno method by sorting, dividing, and distributing these hypernyms around the seed term in the corpus. After that GloVe builds its co-occurrence matrix from the enriched corpus

and builds its word vectors that are used to extract the terms most similar terms to the seed terms from the corpus.

3.3.3 Method 5: GloVe with Iterative Feedback (GloVeIF)

The previous Methods 2, 3 and 4 use WordNet as an auxiliary source of information to attempt to improve GloVe's performance on a smaller training corpus. This method explores another potential source of auxiliary information, the corpus itself, through a process of iterative feedback. In this method, GloVe trains its model, and the most similar terms to the non-trivial laymen terms are iteratively fed back to GloVe to boost the frequencies in the co-occurrence matrix as though there were additional contexts available.

In addition to GloVe parameters, this method requires two more parameters. The first parameter controls the number of top *n* candidate terms to be fed back to GloVe. The second parameter sets the number of iterations to be considered until the system stops iterating. We reported the best setting of these parameters in our experiments. Figure 3.5 shows the architecture of this method.

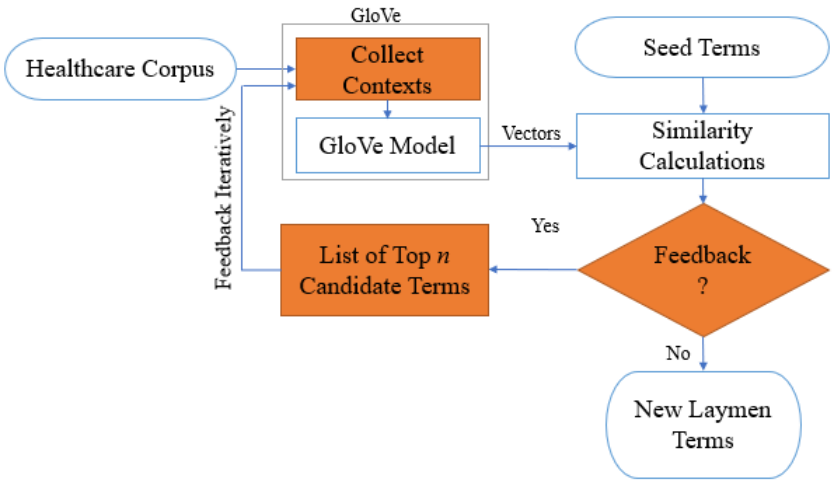

Figure 3. 5 GloVeIF system architecture

GloVe processes the healthcare corpus and collects its co-occurrence contexts. Then, it runs its model to create the best word vectors. After that, for every concept's laymen terms listed in the seed term list, *one* random layman term is picked, and a list of its top *n* candidate terms is produced. We repeat this step until all concepts' randomly picked terms have their top *n* candidate terms. For example, if the seed term list has 50 concepts and each concept has five associated laymen terms, then 50 random laymen terms are picked for every concept, and a list of its top *n* candidate terms is created. We save every selected layman term and its candidate terms into a list.

Now we have the list of laymen terms and their top *n* candidate terms. Two main questions need to be answered:

1) How to feedback the candidate terms into the contexts of other terms?
2) What weight to be assigned to those fed back terms in the context of other terms?

The answer to the first question is that we pick the laymen term first and look to that laymen term in the context of each vocabulary word on the co-occurrence matrix. If that laymen term already appears in the context of vocabulary words, then we feed layman's candidate terms to that word. This process is like the dependency concept: if $A \rightarrow B$, and $B \rightarrow C$, then $A \rightarrow B, C$. To express this process mathematically, assume that *A* is a vocabulary word in the healthcare corpus. Assume *B* is the laymen term with $C_B=\{c_1, c_2, c_3,\ldots, c_n\}$, where C_B is a set of top *n* candidate terms for the laymen term *B*. Assume that the *co-occur* represents the co-occurrence score between *A* and *B*. If *co-occur(A,B)>0,* then *B* occurs in the context of *A*. If this is the case, then C_B should occur in the context of *A*. This is how we feedback candidate terms into the context of other terms.

The answer to the second question is as follows. Some of the fed back candidate terms have co-occurred in the context of vocabulary words while others did not. If the fed back term has a *co-occur* score with the candidate term, then we boost its existing weight with a factor *f*. If it does not, we assign it *f* only. We evaluated a variety of equations for *f*:

$$f = co - occur(A, B). \quad (3.6)$$

$f = cos\_sim(B, C_i)$, where C_i is the current top candidate term. (3.7)

$$f = co - occur(A, B) * cos\_sim(B, C_i) * \sigma, where\ 0 < \sigma \leq 1. \quad (3.8)$$

The f in Equation (3.6) will boost the occurrence of *A* with C_i by assigning it the weight of co-occurrence of A with B while f in Equation (3.7) assigns the score of the cosine similarity between the seed term B and its candidate term C_i. Equation (3.8) is a combination between Equation (3.6) and Equation (3.7) with σ. The σ is a tuning variable that controls the value of *f*. This σ keeps *f* score small comparing to the value of co-occurrence score between *A* and *B*. Having *f* ready, then the new weight, which is the *co-occur(A,Ci)*, equals:

$$\text{co} - \text{occur}(A, C_i) = \begin{cases} \text{co} - \text{occur}(A, C_i) + f, & \text{co} - \text{occur}(A, C_i) > 0 \\ f, & \text{co} - \text{occur}(A, C_i) = 0 \end{cases} \quad (3.9)$$

Here is an example that explains the process of iterative feedback and vector enhancement. Let *B* be the laymen term *pompholyx*, which is a skin disease. *B* has these three top candidate terms: {*dyshidrotic 0.781*, pruritus *0.508, eczema 0.465*}. The score after each candidate term represents the degree of similarity, cos_sim, between *B* and that candidate term. Let *A* be the vocabulary term *vesicle.* Assume that *B* occurs with *A*, and its candidate terms need to be fed into *A* vector, $\overrightarrow{V_A}$. Assume that $\overrightarrow{V_A}$ has the following vocabularies:

	dizzy	pompholyx	dyshidrotic	had	pruritus	for	eczema	irritate	skin
$\overrightarrow{V_A}$	5	15	5	10	1	0	0	0	50

We can see that *A* co-occurred with *B* fifteen times. Also, some of *B* candidate terms co-occurred with *A,* like *dyshidrotic* and *pruritus*, while the candidate term *eczema* does not. According to *f* in equation (3.6), $\overrightarrow{V_A}$ will be:

	dizzy	*pompholyx*	*dyshidrotic*	had	*pruritus*	for	*eczema*	irritate	skin
$\overrightarrow{V_A}$	5	15	20	10	16	0	15	0	50

We can see how the candidate terms boosted to have higher scores. The candidate term *eczema* was zero and now it is 15. If *f* equal equation (3.7), then $\overrightarrow{V_A}$ will be:

	dizzy	*pompholyx*	*dyshidrotic*	had	*pruritus*	for	*eczema*	irritate	skin
$\overrightarrow{V_A}$	5	15	5.781	10	1.508	0	0.465	0	50

Now, if *f* equals equation (3.8), which is the combination between equation (3.6) and equation (3.7) with a tuning parameter σ. Assume that $\sigma = 0.25$, then $\overrightarrow{V_A}$ will be:

	dizzy	*pompholyx*	*dyshidrot*	had	*prurit*	for	*eczema*	irritate	skin
$\overrightarrow{V_A}$	5	15	7.92	10	2.9	0	1.74	0	50

These are all possible cases of *f*. We have tested all *f* settings and reported the results in our next chapter.

3.4 Goal 2: Expanding Professional Medical Concepts

The UMLS has more than 3,800,000 biomedicine concepts, but only 56,000 of them have associated laymen terms, which represents only 1.5% of the UMLS concepts. For our second goal, this is a motivation that we still have about 98% of the UMLS concepts do not have associated laymen terms and not included in the laymen vocabularies. The approach used to enrich concepts revolves around finding synonyms for existing laymen terms (goal 1). However, for the majority of the UMLS concepts, this obviously does not work since there are no existing laymen terms to use as seeds. Thus, the goal here is to identify new laymen terms to add to those professional medical concepts on the UMLS that do not have any associated laymen terms and adding these concepts and their laymen terms to the laymen vocabularies. For this goal, we need

a list of seed terms, but this time the seed terms are a list of professional medical concepts and not laymen terms. This goal should answer two main questions:

1) What are the professional medical concepts out of the three million UMLS concepts to be studied (excluding the existing laymen term)?
2) What source of data will be used to find new laymen terms? Not all concepts would be discussed in an open forum.

Before answering these questions, we have some challenges that need to be addressed. Some of the three million UMLS concepts may have no related laymen terms in an open forum. Some of the UMLS concepts belong to general healthcare categories such as *Devices*, *Geographic Areas, Living Beings*. For example, the UMLS concepts in the *Living Beings* category could have concepts related to the *Amphibian* or *Reptile*. Therefore, it is hard to find a laymen dataset in the field of healthcare that could cover such professional medical concepts.

To narrow down this problem, we used the semantic network that UMLS provides for its database. This network consists of more than fifty semantic types that categorize UMLS concepts into broad topics. To narrow our focus, we studied the current UMLS concepts that have laymen terms to determine which categories they fell in. Our first step to collect seed terms is to select UMLS categories that cover the existing associated laymen terms, and next step is to create a list of seed terms from these categories. Motivated by these reasons, we did an informal experiment and found that most of the currently associated laymen terms fell in the following UMLS categories: *Pharmacologic Substance, Organic Chemical,* and *Disease or Syndrome* categories. It is intuitively logical that laymen terms in these drug and disease categories were added first because people are most likely to discuss their adverse drug reactions and current conditions.

These three categories cover about 460,000 UMLS concepts, and out of these 460,000 concepts, there are 22,826 professional medical concepts have their associate laymen terms. This means we have about 440,000 concepts do not have any associate laymen terms. Now, instead of going over the 3800,000 concepts, we pick seed terms only from the 440,000 concepts. We pick a specific number of concepts from every category and add them to our system seed term list. Figure 3.6 shows the general architecture of this goal.

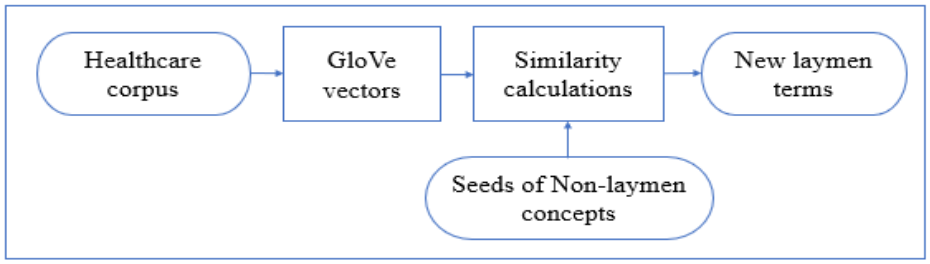

Figure 3. 6 Second goal architecture

3.4.1 Building Healthcare Corpus

This section tackles the second question, which is what is the source of data from which we should get new laymen terms. Such source of data must have good coverage of professional medical concepts. We investigated whether or not we could use the same dataset that we downloaded from *Medhelp.org* website (See section 3.1.1). However, the seed terms we will use for goal 2 will be the UMLS professional names themselves. The issue here is that lay people are less likely to use the professional medical terminology, so they may not appear frequently enough in the corpus to use as seeds since, GloVe might not have enough information available to build accurate word vectors. In case the *Medhelp.org* dataset did not have enough covering of the professional medical concepts, we will collect different datasets from different healthcare social media such as Inspire.com, Yahoo! Answers, and Google search engine.

3.4.2 Goal 2 Methods

In goal two, we will not re-evaluate GloVe and all its enhancements for this goal. Rather, we will test and evaluate this goal using the best method identified during the experimental evaluation of the methods for Goal 1. This is the method that will shows the best performance in term of finding new laymen terms. The difference in Goal 2 is that, rather than using an associated layman's term as the seed, we will use professional terms extracted from the name of the concept itself. We report the results of this goal in the next chapter.

4 Overview

We need to demonstrate the effectiveness of our approach to each of our two main goals: 1) enriching vocabulary for existing laymen terms; and 2) expanding the non-laymen professional medical concepts by adding new laymen terms. To do that, we need a dataset that is a source of ground truth for the laymen terms to be learned and a corpus from which laymen relationships can be learned. These will then be used in an objective experiment to determine which of our methods is most effective at learning "new" vocabulary by comparing the learned words to words in the existing laymen vocabularies. For our first goal, we need a ground truth dataset that has a professional medical concept with at least two associated laymen terms; one used for training and the other(s) for testing. For our second goal, we need a professional medical concept with at least one layman term since we use the professional medical concept description for training and its laymen term(s) for testing. As our baseline for comparison, similar to [140], we use GloVe as our first method. In the next sections, we explain the corpus used for text mining, the source of the ground truth datasets and the metrics used to evaluate the proposed algorithms. The sections after listing all experiments and their results we did to achieve our goals.

4.1 Corpus

For our experiment, we need a large collection of layperson discussions related to the concepts in our laymen vocabularies. Because it hosts wide variety healthcare communities who actively discuss health topics, we have chosen to use *Medhelp.org* as our source of text from which to build our corpus. To select the communities to include in our dataset, we did an informal experiment to find the occurrences of laymen terms on *Medhelp.org* communities. We used the existing laymen terms in the OAC CHV because this vocabulary has good coverage of laymen terms on the UMLS. We found the highest density of these laymen terms occur in communities such as Pregnancy, Women's Health, Neurology, Addiction, Hepatitis-C, Heart Disease, Gastroenterology, Dermatology, and Sexually Transmitted Diseases and Infections (STDs / STIs) communities. We downloaded all the questions on these communities to April 20, 2019. The dataset size is roughly 1.3 Gb and contains approximately 135,000,000 tokens. This dataset is the source of text from which our algorithms attempt to learn new laymen for medical concepts. Table 4.1 shows the downloaded communities with their statistics.

Table 4. 1 Medhelp.org communities' statistics

No.	Community	Posts	Tokens
1.	Addiction	82,488	32,871,561
2.	Pregnancy	308,677	33,989,647
3.	Hepatitis-C	46,894	21,142,999
4.	Neurology	62,061	9,394,044
5.	Dermatology	67,109	8,615,484
6.	STDs / STIs	59,774	7,275,289
8.	Gastroenterology	43,394	6,322,356
9.	Women health	66,336	5,871,323
10.	Heart Disease	33,442	5,735,739
11.	Eye Care	31,283	4,281,328
Total		801,458	135,499,770

Every post on a *Medhelp.org* community has two parts: a question and its answers. For each collected question, we downloaded the question and all its related answers. We also

included the title of the question as part of the question text. We treated each question and its answers as one document because a study in [168] on a healthcare dataset showed that there is no difference between the question and its answer in term of its effects on extracting new terms. We preprocessed each document to remove its stopwords, punctuations, and digits. We also downcased and stemmed all words using the Porter stemmer [164]. The final size of our whole corpus after cleaning and stemming is 865MB.

4.2 Ground Truth Dataset (Seed Terms)

We created the ground truth datasets from the OAC CHV and MedlinePlus vocabularies. For the OAC CHV vocabulary, 43,475 professional medical concepts along with their associated laymen terms were created, and for the MedlinePlus vocabulary, about 1,615 professional medical concepts along with their associated laymen terms were created. For more detail about how we built those datasets, see Section 3.2.1

Because the GloVe embeddings handle only single word vectors, we chose professional medical concepts that have a unigram form, such as flu, fever, fatigue, and swelling. In many cases, the professional medical concept on these two vocabularies has associate laymen terms that have the same names as the concept's name except different morphological forms, such the plural 's', uppercase/lowercase of letters, punctuations, or numbers. We treated these cases and removed any common medical words. After that, we stemmed the terms and listed only the unique terms. For example, the professional medical concept *Tiredness* has the laymen terms *fatigue*, *fatigues*, *fatigued* and *fatiguing*. After stemming, only the term 'fatigu' was kept. To focus on terms for which sufficient contextual data was available, we kept only these laymen terms that occur in the corpus more than 100 times.

To evaluate our system, we need at least two terms for every medical concept. One term is used as the seed terms and we evaluate our algorithms based on their ability to recommend the other term(s) used as a target. Thus, we kept only those medical concepts that have at least two related terms. From the two vocabularies, we created an OAC CHV ground truth dataset of size 944 medical concepts with 2103 seed terms and a MedlinePlus ground truth dataset of size 101 medical concepts with 227 seed terms. Table 4.2 shows an example of some professional medical concepts and their associated laymen terms from the MedlinePlus dataset.

Table 4. 2 Professional medical concepts and unigram associated laymen terms (MedlinePlus dataset)

CUI	Prof. concept	Concept's associated laymen terms		
C0043246	laceration	lacerate	torn	tear
C0015672	fatigue	weariness	tired	fatigued
C0021400	influenza	flu	influenza	grippe

The OAC CHV dataset is nine times bigger than the MedlinePlus dataset (see Figure 4.1a). The reason behind that is that the OAC CHV vocabulary covers 56,000 of the UMLS concepts, whereas the MedlinePlus covers only 2,112 UMLS concepts. Although it is smaller, MedlinePlus represents the future of laymen terms because the NLM updates this resource annually. In contrast, the last update to the OAC CHV was in 2011. Figure 4.1b shows that 37% of MedlinePlus's 101 concepts also appear in the OAC CHV dataset and share the same concepts and laymen terms. This indicates that the OAC CHV is still a good source of laymen terms.

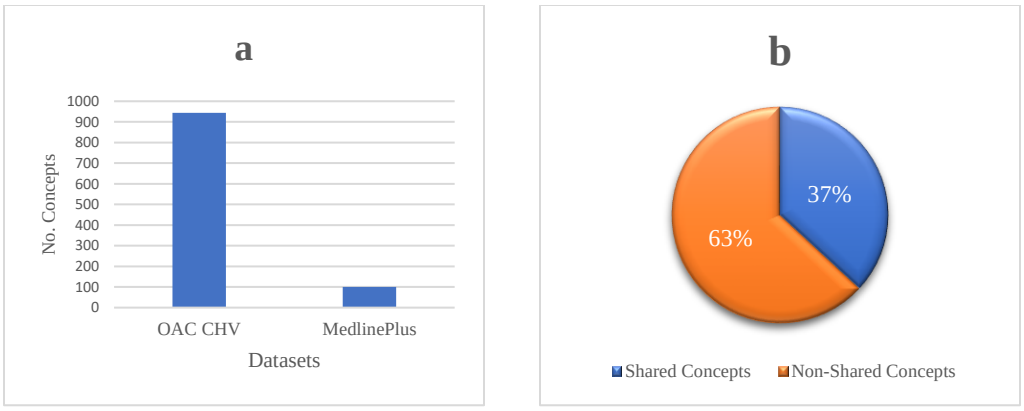

Figure 4. 1 a. Size of the OAC CHV dataset to the MedlinePlus dataset. b. Shared professional concepts and their laymen terms between the MedlinePlus and OAC CHV datasets.

4.3 Evaluation Metrics

We consider GloVe results as the baseline for comparison with the WordNet and iterative feedback algorithms. We evaluate our approach using precision (P), recall (R), and F-score (F) metrics. The F-score represents the harmonic mean of the previous two metrics [169]. We also include the number of concepts (NumCon) that the system could find one or more of its laymen terms. Moreover, we include the Mean Reciprocal Rank (MRR) [170] that measures the rank of the first most similar candidate term in the candidate list. It has a value between 0 and 1, and the closer the MRR to 1, the closer the candidate term position in the candidate list.

Based on a set of professional medical concepts for which we have a seed term and at least one manually identified synonym, we can measure the precision, recall, and F-Score metrics. We measure that according to two criteria: (1) the number of concepts for which the system was able to find at least one synonym; and (2) the total number of synonyms for seed terms the system was able to find across all concepts. We call the metrics used to measure these two criteria the macro and micro average metrics, respectively. The macro average measures the number of the concepts for which the algorithm found a match to the ground truth dataset, while

the micro average measures the number of new laymen terms found. The micro and macro precision, recall, and F-score are computed according to these equations:

$$P_{micro} = \frac{\#\ of\ true\ synonyms\ in\ the\ candidate\ lists}{total\ \#\ of\ terms\ in\ the\ candidate\ lists}, \quad (4.1)$$

$$R_{micro} = \frac{\#\ of\ true\ synonyms\ in\ the\ candidate\ lists}{total\ \#\ of\ synonms\ in\ the\ ground\ truth\ dataset}, \quad (4.2)$$

$$P_{macro} = \frac{\#\ of\ concepts\ whose\ candidate\ list\ contains\ a\ true\ synonym}{total\ \#\ of\ concepts}, \quad (4.3)$$

$$R_{macro} = \frac{\#\ of\ concepts\ whose\ candidate\ list\ contains\ a\ true\ synonym}{total\ \#\ of\ concepts\ in\ the\ ground\ truth\ dataset}, \quad (4.4)$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall}, \quad (4.5)$$

We illustrate these measurements in the following example. Suppose we have a ground truth dataset of size 25 concepts, and every concept has four synonyms terms, which makes 100 synonyms. For every concept, a random synonym term is selected to be a seed term. The remaining 75 synonyms will be used for evaluation. Suppose the algorithm retrieves five candidate terms for each seed term and it is able to generate results for 20 of the seed terms, creating 20 candidate term lists. That makes 100 candidate terms in total. Assume that only 15 out of the 20 candidate lists contain a true synonym, and each list of those 15 lists includes two true synonyms. Thus, this algorithm extracted 30 true laymen terms. Having all this information, then the P_{micro} = 30/100, R_{micro} = 30/75, P_{macro} = 15/20, and R_{macro} = 15/25.

4.4 Goal One: Enriching Existing Laymen Terms

In this goal, we enrich a layman concept with additional vocabulary terms using word embedding algorithms. To do this, we randomly select, for each concept in our truth ontology, one layman term to use as a seed. We compare GloVe with its enhancements with respect to its ability to suggest terms related to the seed that are actual entries in that concept's terminology.

Before comparing GloVe and its enhancements, some experiments need to be done. First, we need to decide whether a domain-specific corpus is better than a general domain corpus or not. Second, GloVe enhancements have many parameters and settings that should be tuned to their best setting before comparing with GloVe. Finally, to have a fair comparison between GloVe and its enhancements, we need to find the best setting for GloVe and apply that best setting to all GloVe's enhancements. After that, we compare GloVe and all its enhancements according to that setting. The next sections explain all these experiments.

4.4.1 Experiment 1: General Corpus vs Domain-Specific Corpus

In this experiment, we investigate what is better to use: a general domain corpus (e.g., Wikipedia) or a corpus collected from a healthcare domain. For the domain-specific corpus, we already have our healthcare corpus that has 135 million tokens. For the general domain corpus, we downloaded the GloVe 42B pre-trained vectors reported in [171]. These are the vectors that GloVe generated from a corpus with 42 billion tokens collected from the web by the Common Crawl organization [172]. GloVe pre-trained these vectors with 300 vector size and 10 window size. We applied GloVe for the same setting, 300 vector size and 10 window size, over our healthcare corpus. We evaluated the two pre-trained vectors using our two ground truth datasets. Table 4.3 illustrates the macro results between these two domains.

Table 4. 3 Comparison between general corpus versus domain-specific corpus over our two ground truth datasets.

		Macro		
	NumCon	**P**	**R**	**F**
OAC CHV				
GloVe 42B	270	34.57	28.6	31.3
GloVe 135M	**432**	**45.76**	**45.76**	**45.76**
MedlinePlus				
GloVe 42B	25	32.05	24.75	27.93
GloVe 135M	**49**	**52.13**	**48.51**	**50.26**
Average				
GloVe 42B	147	33.31	26.675	29.615
GloVe 135M	**240**	**48.945**	**47.135**	**48.01**

We can see from Table 4.3 that even though our corpus is 0.3% of the general corpus (42 billion tokens), our corpus provides more laymen terms than the general one. Over the two ground truth datasets, our corpus provided 48.01% of F-score compering to 29.6% F-score from the 42B corpus. That confirms that using a domain-specific corpus would be better than using a general domain corpus.

4.4.2 Experiment 2: Tuning GloVe with WordNet Enhancement

The WordNet ontology provides our corpus with extra information about the laymen terms that resided in the corpus. This auxiliary source can be used in many ways. This section explains some experiments that we did to decide the better ways to use the WordNet for our laymen terms enrichment goal.

From our WordNet investigations, we found that the synonyms, hypernyms, and hyponyms relations represent a good source of information that can be used to enrich our healthcare corpus. We call a list of words coming from these WordNet relations as the relational synsets. Here is how we used these relational synsets: When a seed term occurs in the corpus, a list of its relational synsets is extracted from the WordNet. This list of synsets is sorted according to the synset's degree of similarity to the seed term. Then, it is divided into two lists and distributed around the context of the seed term. See section 3.3.2 for more details about how we used the WordNet ontology.

In another experiment, we investigated whether adding synsets around the seed term in a repeated way would improve GloVe results or not. For example, if a seed term *X* occurs 5000 times in the corpus and the list of its WordNet synsets has 10 synsets. Then, enriching *X* with these synsets would lead to 50,000 words around *X*. The question here is: would that help improve GloVe results or add more noise and lead to bad outcomes?

We found that having a cut-off to limit the number of times to enrich the seed term improves GloVe results. For this improvement, we used our corpus that was enriched with the WordNet synonyms. We call GloVe with this relation as GloVeSyno as a short for GloVe with WordNet Synonyms relations. We call the parameter that controls the number of times to enrich a seed term by *MaxEnrich*. We implemented GloVeSyno with a vector size of 100 and a window size of 10. The *MaxEnrich* is set between 1000 and 15,000, adding 1000 incrementally. Figure 4.2 a and Figure 4.3 show different settings for the *MaxEnrich* parameter over the two ground truth datasets. From these figures, we can see that over the two ground truth datasets, setting *MaxEnrich to* 3000 improved GloVeSyno results. We used that best setting for all WordNet relational synsets methods.

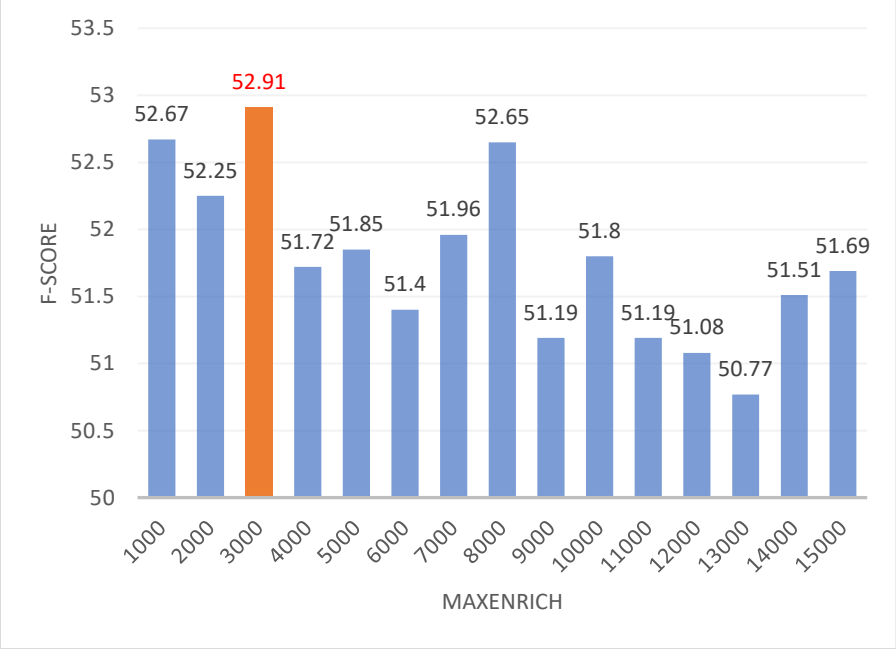

Figure 4. 2 F-score results of GloVeSyno over the OAC CHV dataset with different context enrichment settings.

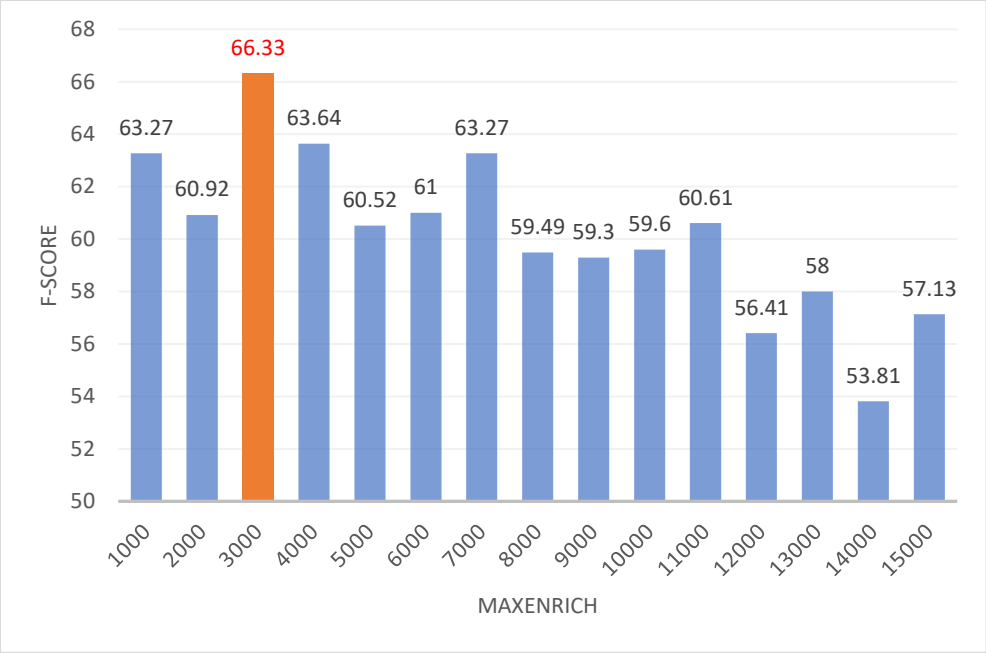

Figure 4. 3 F-score results of GloVeSyno over the MedlinePlus dataset with different context enrichment settings.

4.4.3 Experiment 3: Tuning GloVeIF

After tuning GloVe with WordNet methods to their best setting, it is time to set the GloVe with Iterative Feedback (GloVeIF) method to its best setting. First, we need to see which weighting method is the best for GloVeIF. We have reported three weighting equations to boost the occurrence of the fed back candidate terms in the GloVe co-occurrence matrix (See Section 3.3.3). All our experiments for the GloVeIF algorithm will be evaluated using the OAC CHV dataset only.

4.4.3.1 Tuning σ Equation 3.8

Before comparing the three weighting methods, we must investigate the most effective value for a tuning parameter (See Equation (3.8) σ, where $0 < \sigma \leq 1$. σ keeps f score small comparing to the value of co-occurrence score between *vocabulary word* and *layman term*. To find the best setting for that parameter, we run the GloVeIF over the whole corpus with a vector size of 100, a

window size of 10, one-time iterative feedback, and 20 candidate terms for every seed term. For σ, we picked numbers between 0 and 1. Figure 4.4 shows the results of this experiment.

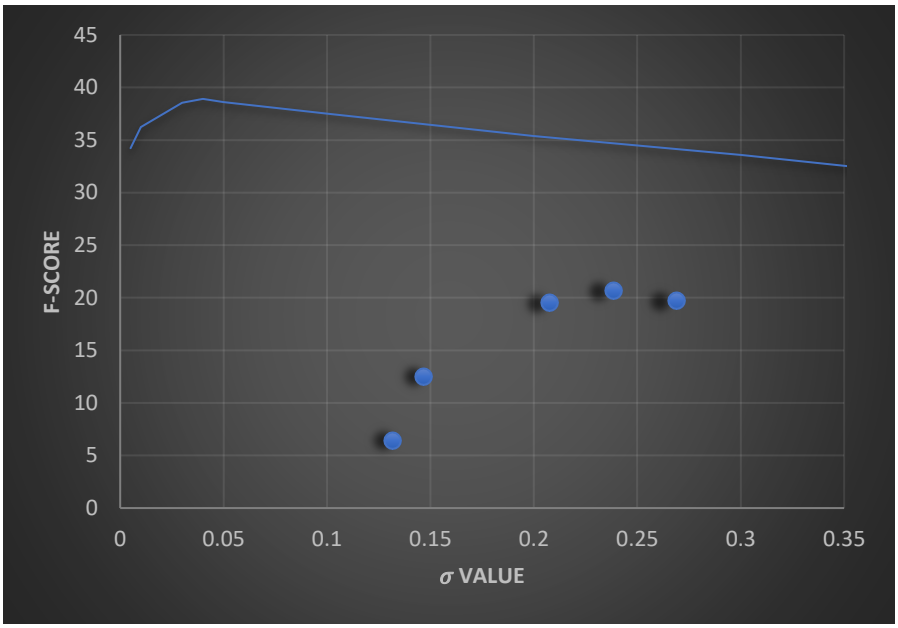

Figure 4. 4 F-score results for tuning σ to find the best setting for Equation (3.8).

We can see from Figure 4.4 that the F-score was at its best performance at σ = 0.04, thus we will use σ = 0.04 in all our next experiments to compare between the other weighting methods.

4.4.3.2 Finding the Best Weighting Method for GloVeIF

Now we compare between all the presented weighting methods we reported in Chapter 3 (See Equation (3.6) that is based on co-occurrences, Equation (3.7) that is based on context similarity, and Equation (3.8) which is a combination of the previous two factors). We already have the best setting for Equation (3.8) reported in the previous section. For the other weighting methods, we run GloVeIF with the same setting we used in the previous section, 100 vector size and 10 window size. Figure 4.5 shows the F-score results over different weighting methods.

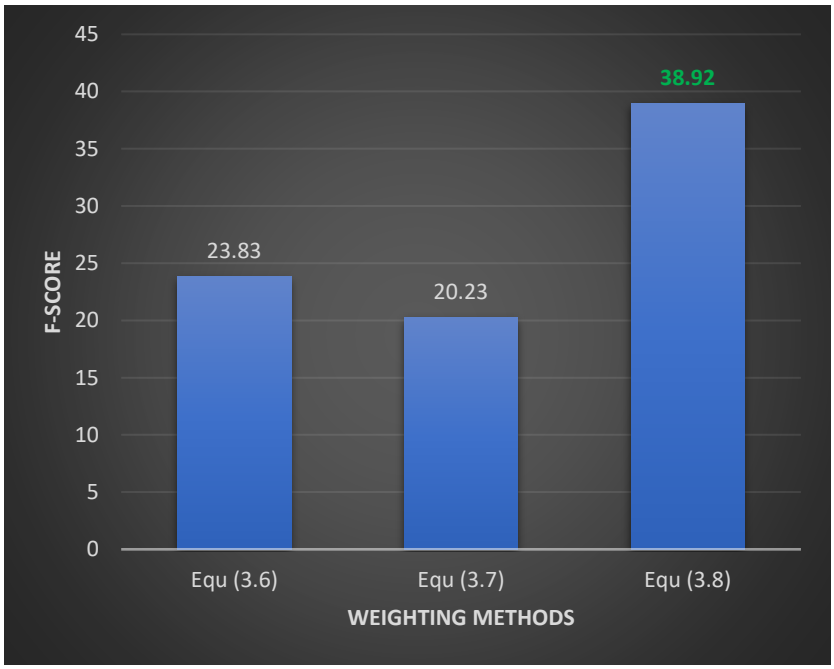

Figure 4. 5 F-score results for different weighting methods.

We can see from Figure 4.5 that the highest F-score reported was with the weighting method for Equation (3.8) with an F-score of 38.92%. This method will be used in all next GloVeIF experiments.

4.4.3.3 How Many Candidate Terms to Iteratively Feedback to GloVeIF?

In our previous GloVeIF experiments, we set the number of candidate terms to be fed back to GloVeIF to 20 candidate terms. Although that has worked well for us, we next want to identify the best number of candidate terms to use. For GloVeIF. We created different lists of candidate terms start from 5 and ends at 30 with 5 terms increase. Figure 4.6 shows the F-score results of GloVeIF over these different candidate terms lists.

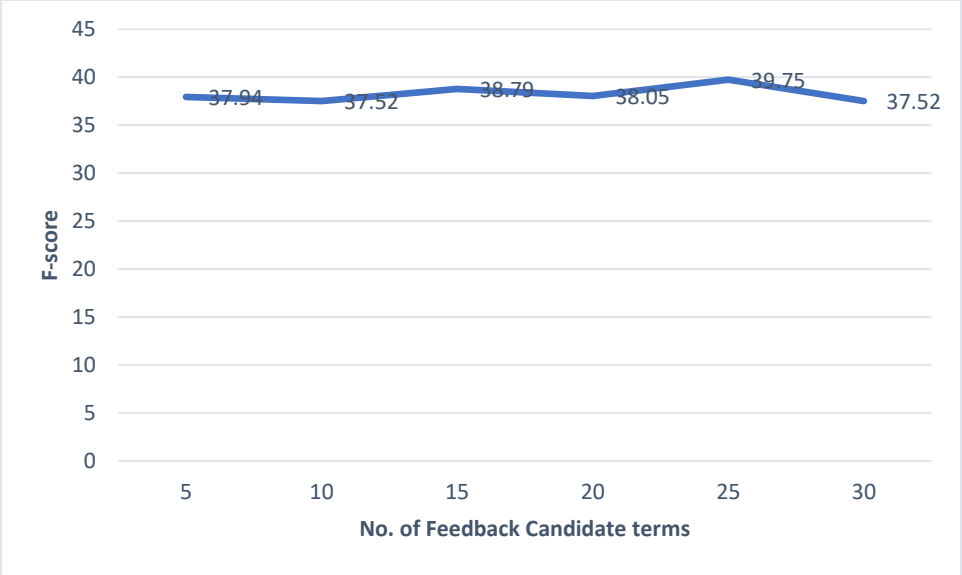

Figure 4. 6 F-score results for different size feedback candidate terms.

We can see from Figure 4.6 that the highest F-score reported was with a feedback list of size 25. The F-score value starts to go down when the list size is smaller or bigger than 25. However, if we measure the difference percentage between the maximum and minimum F-score values, we get a 6% difference. This difference shows that most of the feedback candidate lists provided good results, and there is not that much difference between their results. However, for our experiment, we will use the best-performing list size, 25 candidate terms.

4.4.4 Experiment 4: Tuning GloVe to its Best Setting

Now we have GloVe with WordNet ontology and GloVe with iterative feedback set to their best settings. It is time to set GloVe to its best setting so that we can make a fair comparison between GloVe and its enhancements. To tune GloVe to its best setting, we used the larger of our two datasets, the OAC CHV. GloVe has many hyperparameters, but the vector size and the window size parameters have a significant effect on the results. We evaluated GloVe using the 944 professional concepts in this dataset on different vector sizes (*100, 200, 300, 400*), varying the window size (*10, 20, 30, 40*) for each vector size. We set the candidate list size to $n = 10$.

Figure 4.7 shows the macro F-score results of GloVe according to these different vector and window sizes. In general, the F-score results declined with any window size greater than 30. The highest F-score was reported at a vector of size 400 and a window of size 30. Thus, we used these settings for all following experiments.

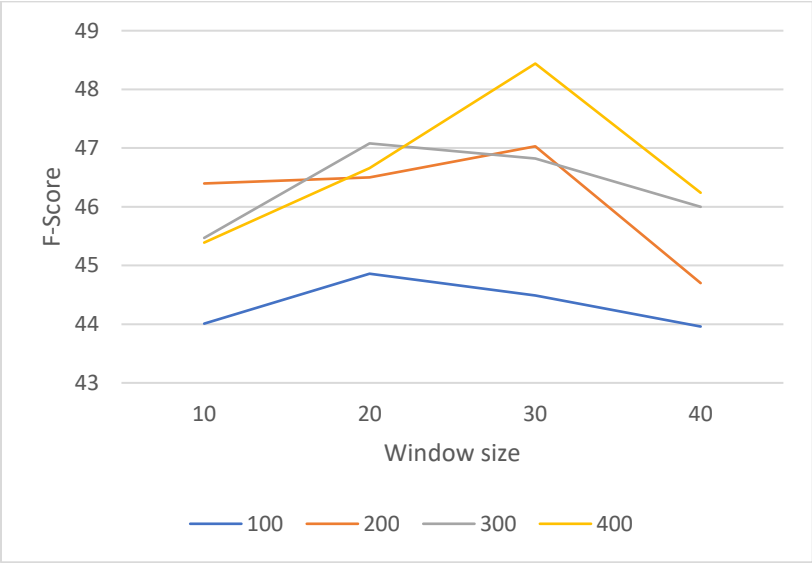

Figure 4. 7 The Macro F-Score for GloVe with Different Vector and Window Sizes.

Table 4.4 reports the micro-precision for GloVe over the same parameter settings. We can see that the micro precision is very low due to the size of the candidate lists created. In particular, we are testing with 944 professional concept seed terms, and the size of the candidate list is set to 10, so we generate 944x10=9440 candidate terms. However, there are only 2103 truth synonyms, to the micro-averages are guaranteed to be quite low. To compensate, we need to determine a good size for the candidate list that balances recall and precision. This is discussed further in Section 4.4.5.1.

Table 4. 4 The micro-precision of GloVe.

		Micro		
Vector Size	**NumCon**	**P**	**R**	**F**
100	420	4.78	38.91	8.51
200	444	5.07	41.33	9
300	442	5.16	42.02	9.19
400	**457**	**5.28**	**42.97**	**9.41**

4.4.5 GloVe Verses its Enhancements Over the Whole Corpus

Using the best GloVe settings reported in the previous experiment, we next evaluate the GloVeSyno, GloVeHypo, GloVeHyper, and GloVeIF algorithms to determine whether or not they can improve on GloVe's ability to find layman terms. GloVe and all its enhancements run with 400 for the vector size and 30 for the window size. Table 4.5 shows a comparison between the results of these algorithms for the OAC CHV and MedlinePlus datasets. The evaluation was done using a candidate list of size n = 10. We report here the macro accuracy of the system for all algorithms, which is based on the number of concepts for which a ground truth result was found.

Table 4. 5 Evaluation of GloVe and its enhancements over the OAC CHV and MedlinePlus datasets on the whole corpus.

		Macro			
	NumCon	**P**	**R**	**F**	**MRR**
OAC CHV					
GloVe	457	48.46	48.41	48.44	0.29
GloVeSyno	***546***	***57.9***	***57.84***	***57.87***	***0.35***
GloVeHypo	280	29.69	29.66	29.68	0.33
GloVeHyper	433	45.92	45.87	45.89	0.35
GloVeIF	389	41.25	41.21	41.23	0.34
MedlinePlus					
GloVe	48	51.06	47.52	49.23	0.38
GloVeSyno	***63***	***66.32***	***62.38***	***64.29***	***0.36***
GloVeHypo	32	33.33	31.68	32.49	0.37
GloVeHyper	35	37.23	34.65	35.9	0.35
GloVeIF	36	39.56	35.64	37.5	0.32

We can see from Table 4.5 that GloVeSyno outperformed the other algorithms. It enriched synonyms to 57% (546) of the professional medical concepts listed in the OAC CHV dataset and more than 62% (63) of the concepts in the MedlinePlus dataset. Table 4.6 presents the algorithms' performance averaged over the two datasets. On average, the GloVeSyno algorithm produced an F-score relative improvement of 25% comparing to the GloVe. Moreover, the GloVeSyno reported the highest MRR over all the other algorithms, which shows that the first most similar candidate term to the seed term fell approximately in the 2nd position of the candidate list. A comparison between GloVe and GloVeSyno results over a whole corpus with on the OAC CHV dataset showed that GloVeSyno results were statistically significant with p-value = 2.73993E-12, which is almost a zero. GloVeSyno also showed a high significance over the MedlinePlus dataset with p-value=0.000287315.

Table 4. 6 The average results of GloVe and its enhancements over the OAC CHV and MedlinePlus datasets on the whole corpus.

		Macro				
Algorithm	**NumCon**	**P**	**R**	**F**	**MRR**	**F-score Rel-Improv.**
Basic GloVe	252.5	49.76	47.965	48.835	0.335	
GloVeSyno	***304.5***	***62.11***	***60.11***	***61.08***	***0.355***	***25%***
GloVeHypo	156	31.51	30.67	31.085	0.350	-36%
GloVeHyper	234	41.575	40.26	40.895	0.350	-16%
GloVeIF	212	40.405	38.425	39.365	0.33	-19%

The results of GloVeHypo, GloVeHyper, and GloVeIF were not good comparing to the other algorithms. The reason is that the hyponyms provide a very specific layman term synsets. For example, the hyponyms of the laymen term *edema* are *angioedema*, *atrophedema*, *giant hives*, *periodic edema*, *Quincke'*, *papilledema*, and *anasarca*. Such hypos are specific names of the laymen term *edema*, and they might not be listed in ground truth datasets. We believe that the

GloVeHypo algorithm results are promising, but a more generalized and bigger size ground truth dataset is required to prove that.

On the other hand, the GloVeHyper algorithm was not good comparing to GloVe. However, it is better than the GloVeHypo algorithm. This algorithm did not get a good result due to the degree of abstraction that the hypernym relations provide. For example, the hypernym *contagious_disease* represents many laymen terms, such as *flu, rubeola,* and *scarlatina*. Having such hypernym in the context of a layman term did not lead to good results. The hypernym *contagious_disease* is a very general relation that can represent different kind of diseases.

GloVeIF also showed not good results comparing to GloVe. This indicates that feeding candidate terms iteratively to GloVe co-occurrence matrix added noise. However, GloVeIF showed a better performance with small size corpus (See Section 4.4.3.4). To those who have a small size corpus, GloVeIF is a good algorithm to try to find new laymen terms.

To illustrate the effectiveness of GloVeSyno, we show a seed term the candidate synonyms for a selection of concepts in Table 4.7. The candidate synonyms that appear in the ground truth list of synonyms are shown shaded. Although only 14 true synonyms from 7 concepts were found, we note that many of candidate synonyms seem to be good matches even though they do not appear in the official laymen vocabularies. These results are promising and could be used to enrich medical concepts with missing laymen terms. They could also be used by healthcare retrieval systems to direct laypersons to the correct healthcare topic.

Table 4. 7 Sample of the GloVeSyno output.

CUI	laymen term	Candidate synonyms						
C0015967	**feverish**	febric	febril	pyrexia	fever	chili_pepp	chilli	influenza
C0020505	**overeat**	gormand	pig_out	ingurgit	gormandis	scarf_out	overindulg	gourmand
C0013604	**edema**	oedema	hydrop	dropsi	swell	puffi	ascit	crestless
C0039070	**syncop**	swoon	deliquium	faint	vasovag	neurocardi	dizzi	lighthead
C0015726	**fear**	fright	afraid	scare	terrifi	scari	panic	anxieti
C0014544	**seizur**	ictus	seiz	raptus	prehend	shanghaier	seizer	clutch
C0036916	**stds**	std	gonorrhea	encount	chlamydia	hiv	herp	syphili

4.4.5.1 Improving GloVeSyno Micro Accuracy

From our previous experiment, we conclude that the GloVeSyno algorithm was the most effective algorithm. However, we explore it in more detail to see if we can improve its accuracy by selecting an appropriate number of candidate synonyms from the candidate lists. We report evaluation results according to the ground truth datasets, OAC CHV and MedlinePlus. We varied the number of synonyms selected from the candidate lists n=1 to n=100 and measured the micro recall, precision, and F-score. Figure 4.8 shows the F-score results and the number of concepts for which at least one true synonym was extracted. This figure reports the results of the GloVeSyno algorithm over the OAC CHV dataset. The F-score is maximized with n=3 with an F-score of 19.06% and 365 out of 944 concepts enriched. After that, it starts to decline quickly and at n=20 the F-score is only 6.75%, which further declines to 1.7% at n=100. We note that the number of concepts affected rose quickly until n=7, but then grows more slowly. The best results are with n=2 with an F-score of 19.11%. At this setting, 287 of the 944 concepts are enriched with a micro-precision of 15.43% and recall of 25.11%.

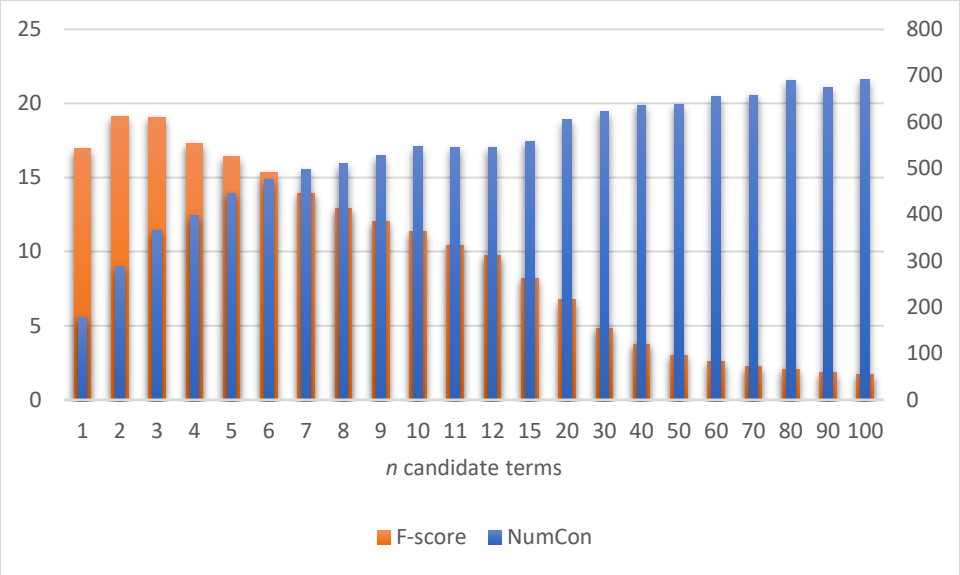

Figure 4. 8 Micro F-Score and the number of concepts for the GloVeSyno algorithm over the OAC CHV dataset.

The evaluation results with the MedlinePlus dataset are similar to those reported for the OAC CHV dataset (See Figure 4.9). The F-score was at its highest score at *n*=2 with an F-score of 23.12% and 33 out of 101 concepts enriched. The F-score decreased quickly at *n*= 30 and was at its lowest score at *n*=100 with an F-score of 1.81%. The number of enriched concepts grew quickly until *n*=6 and stabilized after *n*=9 between 64 and 74 enriched concepts.

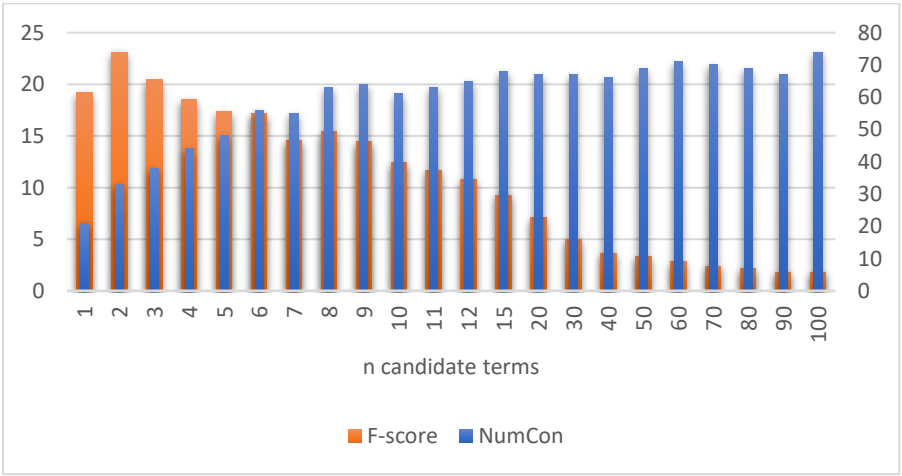

Figure 4. 9 Micro F-Score and the number of concepts for the GloVeSyno algorithm over the MedlinePlus dataset.

Over the two datasets, the best results are with *n*=2. Figure 4.10 shows the F-score over the Precision and recall for the two datasets. Despite the difference in the number of concepts between the two ground truth datasets, the results show that the F-score is the best at *n*=2. The figure shows that the behaviors of the GloVeSyno over the two datasets are almost the same over different candidate list settings.

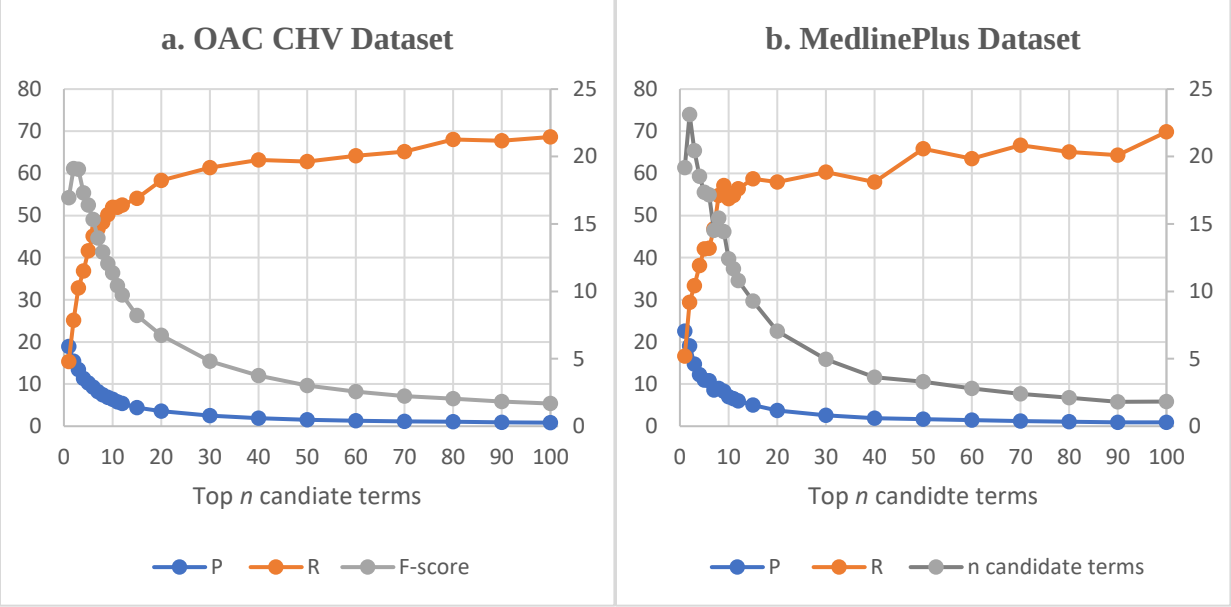

Figure 4. 10 F-Score results over the Precision and Recall for the GloVeSyno algorithm over the OAC CHV and MedlinePlus datasets

4.4.6 GloVe Verses its Enhancements Over a Small Size Corpus

Using our whole corpus of 865MB, we and found that GloVeSyno outperformed all other algorithms. We also found that GloVe was the second-best algorithm. To explore whether or not some of our other algorithms perform better with smaller corpora, we created a smaller subset of our corpus by selecting100MB at random. We then applied all our algorithms on that corpus with the same parameter settings reported in the previous section. Table 4.8 presents the results of this experiment.

Table 4. 8 Evaluation of GloVe and its enhancements over the OAC CHV and MedlinePlus datasets on a corpus of size 100MB.

		Macro		
	NumCon	**P**	**R**	**F**
OAC CHV				
GloVe	284	31.38	30.08	30.72
GloVeSyno	***396***	***46.75***	***41.95***	***44.22***
GloVeHypo	70	9.15	7.42	8.19
GloVeHyper	139	19.52	14.72	16.79
GloVeIF	***312***	***34.48***	***33.05***	***33.75***
MedlinePlus				
GloVe	20	26.67	19.8	22.73
GloVeSyno	***38***	***50.0***	***37.62***	***42.94***
GloVeHypo	6	8.96	5.94	7.14
GloVeHyper	10	16.39	9.9	12.35
GloVeIF	***22***	***29.33***	***21.78***	***25.0***

Table 4.9 presents the algorithms' performance averaged over the two datasets on a 100MB corpus. With small size corpus, GloVeSyno and GloVeIF both outperform GloVe. The concept of feeding candidate terms iteratively to GloVe made a contribution here and improved GloVe results. So, although GloVeIF did not improve our results when there is sufficient training data, when a small corpus is used, it can help boost the performance. Once again, GloVeSyno is the best overall algorithm. This indicates that the related words added to the corpus from WordNet are more accurate than the words added from GloVe's own similar word vectors. On average, the GloVeSyno algorithm produced an F-score relative improvement of 63% over GloVe. GloVeIF provided a much smaller 9.91% relative improvement.

Table 4. 9 The average results of GloVe and its enhancements over the OAC CHV and MedlinePlus datasets on a corpus of size 100MB.

		Macro			
Algorithm	**NumCon**	**P**	**R**	**F**	**F-score Rel-Improv.**
Basic GloVe	152	29.02	24.94	26.72	
GloVeSyno	***217***	***48.37***	***39.78***	***43.58***	***63.06***
GloVeHypo	38	9.05	6.68	7.66	-71.31
GloVeHyper	74.5	17.95	12.31	14.57	-45.48
GloVeIF	***167***	***31.90***	***27.41***	***29.37***	***9.91***

We next tested our results using Student's t-test to see if these results were statistically significant. GloVeIF's improvement over GloVe on the MedlinePlus dataset was not statistically significant ($p = 0.22$) but it was on the OAC CHV dataset ($p=0.0038$). In contrast, GloVeSyno was statistically significantly better than GloVe on both datasets, MedlinePlus ($p = 0.0076$) and OAC CHV ($p = 1.41391E-14$).

4.5 Goal Two: Ontology Expansion

The purpose of this experiment is to evaluate our second goal of expanding the number of professional concepts who have associated laymen's terms. This goal is different from our first goal in that it deals with the description of the medical concept itself. So, the seeds in this goal are not laymen terms but professional medical concepts.

4.5.1 Corpus

As in Section 3.4, this goal faces the problem of finding a source of text that has good coverage of professional medical concepts. Although primarily focused on informal discussions, our previous investigations showed that the MedHelp.org corpus that we collected also contains many occurrences of professional medical terms. Therefore, we decided to use this corpus again to evaluate our second goal.

4.5.2 Ground Truth Ontology

In our first goal, the input to the proposed system is a set of existing laymen terms, and the output is a set of new, related laymen terms. In our second goal, the input to the proposed system is a professional medical concept, and the output is a list of new, related laymen terms. To evaluate this, we need a ground truth dataset. Once again, we use the ground truth datasets from our evaluation of our first goal with the modification that we need to treat the concepts in the truth ontology as though they do not have laymen terms and evaluate whether or not our

algorithms can learn the associated laymen terms directly from the medical concept information. From the OAC CHV and MedlinePlus datasets, we picked only concepts with associated laymen's terms that have a unigram form because the GloVe embeddings handle only single word vectors. Although the concepts we are working with for our ground truth have associated laymen's terms, they are not used in by our algorithms for goal 2. Thus, if we demonstrate that we can discover the ground truth laymen's terms for these concepts, we believe that our approach would then be applicable to other medical concepts for which no laymen's terms are available.

Of the 944 professional medical concepts in the OAC CHV dataset, we picked 827 concepts that have a unigram form. Out of 101 professional medical concepts for the MedlinePlus dataset, we picked 71 medical concepts that have a unigram form. Table 4.10 shows some examples of professional medical concepts and their associated laymen terms from the OAC CHV dataset.

Table 4. 10 Professional medical concepts and their associated laymen terms from the OAC CHV dataset.

No.	CUI	Prof. Concepts	Laymen terms			
1	C0027497	nausea	nauseous	nauseant	queasy	
2	C0036658	esthesia	sense	sensate	sensory	
3	C0003123	anorexia	appetite	anorectic	anorexia	anorexia
4	C0033975	psychosis	mental	psychotic		
5	C0031354	pharynx	throat	pharyngeal		
6	C0003578	apnea	breath	apnea		

4.5.3 Metrics and Methods

We evaluated this goal using the same metrics discussed in Section 4.3. For the second goal methods, we applied the best method reported in our first goal experiments. GloVeSyno

outperformed all the other GloVe methods. So, it will be the algorithm that will test and evaluate this goal. GloVeSyno will be compared with GloVe to see the performance of these two algorithms over the two ground truth datasets for our second goal.

4.5.4 Results

We implemented GloVe and GloVeSyno algorithms using the best setting reported previously for GloVe, a 400 vector size, 30 window size, and top 10 candidate terms for evaluation. Table 4.10 illustrates the results of applying these algorithms on the OAC CHV and MedlinePlus datasets. The results show good findings of new laymen terms for different professional medical concepts. GloVeSyno algorithm also outperformed GloVe in this experiment. Table 4.11 also reports the average of the macro accuracy over the two ground truth datasets. On average, there was a 29.2% F-score relative improvement over Glove. The GloVeSyno showed not only a good macro accuracy but also a good MRR. On average, the GloVeSyno algorithm reported an MRR of 0.32 comparing to GloVe, which is 0.28, which makes a 14% relative improvement. A comparison between GloVe and GloVeSyno results over the OAC CHV dataset showed that GloVeSyno results were statistically significant with p-value = 8.4144E-16. It also reported statistically significant results over the MedlinePlus dataset with a p-value = 6.78349E-05.

Table 4. 11 Evaluation of GloVeSyno algorithm over the OAC CHV and MedlinePlus datasets for the 2nd goal.

		Macro				
	NumCon	**P**	**R**	**F**	**MRR**	**F Rel. Improvement**
OAC CHV						
Basic GloVe	354	44.36	42.81	43.57	0.27	
GloVeSyno	**448**	**55.86**	**54.17**	**55**	**0.36**	**26.23**
MedlinePlus						
Basic GloVe	38	55.88	53.52	54.68	0.3	
GloVeSyno	**50**	**73.53**	**70.42**	**71.94**	**0.29**	**31.57**
Average						
Basic GloVe	196	50.12	48.165	49.125	0.285	
GloVeSyno	**249**	**64.695**	**62.295**	**63.47**	**0.325**	**29.2**

The results in Table 4.11 show that the MedHelp.org corpus has good coverage of professional medical concepts. The GloVeSyno algorithm was able to determine many laymen terms from such corpus. This proves that this corpus is a good source to evaluate our 2nd goal.

Table 4.12 shows some examples of the professional medical concepts and their 5 top candidate terms from the GloVeSyno algorithm. We have highlighted those candidate terms that are in the ground truth datasets.

Table 4. 12 Sample of the GloVeSyno output for the 2nd goal.

Prof. Concepts	Top candidate terms				
heartburn	pyrosis	reflux	indigest	upset_stomach	dyspepsia
measles	rubeola	morbilli	mump	epidemic_roseola	german_measl
farsighted	farsight	presbyopia	longsighted	hyperopia	hypermetropy
icterus	jaundice	bilirubin	diabetopaedia	yellow	icter
dyspepsia	upset_stomach	indigest	stomach_upset	dyspept	heartburn
edema	oedema	hydrops	dropsy	swell	puffy
hemorrhage	haemorrhage	shed_blood	bleed	haemophile	hemophiliac
mdma	ecstasy	rapture	methylenedioxy-methamphetamin	ecstatic	lsd
chickenpox	varicella	zoster	herpes	zost	shingle

We can see from Table 4.11 that even if some of the candidate terms are not in the ground truth datasets, they are highly related to their professional medical concept. For example, the professional medical concept *heartburn* has the candidate terms *{reflux, indigest, upset_stomach, dyspepsia}*. These terms are not listed in the ground truth dataset, but these terms are related to the medical concept *heartburn*. The GloVeSyno algorithm reported the highest MRR over GloVe. We can see that from the results in Table 4.11, in which many of the ground truth associated laymen terms have been detected in the first position of the top candidate terms.

GloVeSyno also defined abbreviations. For example, the medical concept *mdma* has the candidate term *methylenedioxymethamphetamine*, which defines that *mdma* abbreviation. Not only that, GloVeSyno provided a related abbreviation to that concept, which is the *lsd* that refers to the *lysergic acid diethylamide*. If a medical concept shares the same meaning with another medical concept such as *heartburn* and *dyspepsia*, we can see that they are sharing many of their candidate terms, such as *indigest*, *upset_stomach*. Some of the candidate terms have a hyphen in their names. This hyphen comes from the synonyms that WordNet ontology provide.

4.5.4.1 Micro Accuracy of GloVeSyno for the 2nd Goal Experiment

The highest micro accuracy reported in our 1st goal was with n=2, where n is the size of the candidate list (See Section 4.4.5.1). We set that best setting for our 2nd goal experiment and reported the results in Table 4.13.

Table 4. 13 Micro Accuracy for the GloVeSyno for the 2nd goal experiment

		Micro			
	NumCon	**P**	**R**	**F**	**F Rel. Improvement**
OAC CHV					
Basic GloVe	214	14.1	20.16	16.59	
GloVeSyno	**253**	**16.33**	**23.48**	**19.26**	**16**
MedlinePlus					
Basic GloVe	18	13.24	19.15	15.65	
GloVeSyno	**27**	**22.79**	**32.98**	**26.96**	**72**
Average					
Basic GloVe	116	50.12	48.165	49.125	
GloVeSyno	**140**	**22.79**	**22.79**	**26.96**	**44**

Although the two algorithms' micro accuracy over the two datasets is low, the GloVeSyno is more accurate than GloVe. On average, GloVeSyno provided a 44% F-score relative improvement comparing GloVe results.

For our second goal, GloVe was able to detect laymen terms for many professional medical concepts. Moreover, GloVe enhancement, GloVeSyno, outperformed GloVe and

provided more laymen terms than GloVe. The results are promising and show word embedding algorithms' ability to be applied in the medical domain and find many formal and informal terms. The results of such algorithms can be applied not only to enrich and expand vocabularies but also in many other domains, such as translating, search engine optimizations, and many others.

5 Conclusion

5.1 Summary

Ontologies play a main role in providing organized and machine-readable data. They have been applied in different domains such as semantic web applications, text simplification, translating, text annotations, and word disambiguation. Creating ontologies manually is time-consuming and requires a lot of human effort. Ontology learning is the new era of building ontologies automatically or semi-automatically. NLP tools, statistics approaches, machine leading, data mining are all methods used to learn ontologies. Ontologies can be constructed using different text resources, such as structured, semi-structured, and unstructured text documents.

An ontology defines different levels of conceptualization with data provided from different vocabularies and terminologies. There are many types of ontologies, and this work focuses on healthcare ontologies. Specifically, the UMLS ontology and its consumer health vocabularies, which are the OAC CHV and MedlinePlus vocabularies. These vocabularies provide easy and straightforward laymen terms that have been mapped to many professional medical concepts. However, many professional medical concepts still miss some of their laymen terms, and many other professional medical concepts do not have any mapped laymen terms.

This research presents an automatic approach to tackle two goals: (1) Enriching existing laymen terms with new laymen terms (2) Expanding non-laymen professional medical concepts with laymen terms. GloVe word embedding and its enhancements are used to achieve these two goals. GloVe enhanced once using an auxiliary lexical source called WordNet, and another enhancement utilizing the concept of iteratively feeding GloVe's candidate terms. The presented approaches were evaluated using a healthcare corpus downloaded from a healthcare social media platform called MedHelp.org. Two standard laymen vocabularies, OAC CHV, and MedlinePlus

were used to test system performance. Only unigram laymen terms and professional medical concepts are picked because GloVe only provides unigram vectors. Given a seed term selected from a concept in the ontology, we measured our algorithms' ability to automatically extract synonyms for those terms that appeared in the ground truth concept.

We used the WordNet ontology to expand the healthcare corpus by including synonyms, hyponyms, and hypernyms for each layman term occurrence in the corpus. We called GloVe with these relations by GloVeSyno, GloVeHypo, and GloVeHyper according to each relation name. The other enhancement to GloVe used the idea of feeding GloVe's candidate terms iteratively back to GloVe to boost their occurrence in the vector of seed terms on the GloVe's co-occurrence matrix. We called this approach by GloVeIF, which is short for GloVe Iterative Feedback. We implemented GloVe and its enhancements to our first goal, and the best approach from our first goal was used to evaluate our second goal.

Before evaluating our first goal, we compared a general domain and domain-specific corpora and found that our healthcare corpus provided better results than the general one. After that, we set GloVe and its enhancement to their best setting. Because some of the approaches showed better performance over a small size corpus, we decided to evaluate our first goal over a small and a whole size corpus. From our whole corpus experiments, GloVe was able to enrich existing laymen terms with new laymen terms with an F-score of 48.44%. GloVeSyno was the best approach overall approaches and outperformed GloVe with an F-score of 61%, making a 25% relative improvement over GloVe. Moreover, a comparison between GloVe and GloVeSyno results over a whole corpus with the OAC CHV dataset showed that GloVeSyno results were statistically significant with a p-value = 2.73993E-12, which is almost a zero.

GloVeSyno also showed a high significance over the MedlinePlus dataset with p-value=0.000287315.

Over a small size corpus, GloVeSyno and GloVeIF outperformed GloVe. The GloVeIF improvement emerged when a small size corpus is applied. However, GloVeSyno was the best over other presented approaches and corpora sizes. GloVeSyno algorithm produced an F-score relative improvement of 63% comparing to the GloVe. GloVeIF provided a 9.91% relative improvement, which is small comparing to the improvement that GloVeSyno provided. Also, GloVeIF provided a p-value = 0.22 over the MedlinePlus dataset, which does not show a significance comparing to GloVeSyno that provided p-value= 0.0076.

Over the OAC CHV dataset, they both provided significant results with a p-value = 0.0038 for GloVeIF and p-value= 1.41391E-14 for GloVeSyno. Over the two datasets, GloVeSyno showed the highest statistical significance and was the best over small and whole corpus experiments.

For our second goal evaluation, we used the best approach reported in the first goal, which is GloVeSyno and compared this approach with Glove. In general, GloVe was able to expand non-laymen professional medical concepts with an F-score of 49.12%. GloVeSyno also outperformed GloVe with an F-score of 63.47%, which makes a 29.2% relative improvement. For the significance of GloVeSyno results, a comparison between GloVe and GloVeSyno results over the OAC CHV dataset showed that GloVeSyno results were statistically significant with a p-value = 8.4144E-16. It also reported statistically significant results over the MedlinePlus dataset with a p-value = 6.78349E-05.

To sum up, for our two goals, GloVe enriched and expanded laymen vocabularies from our healthcare corpus. GloVeSyno outperformed all other algorithms and provided good

enrichment and enhancements results. All GloVeSyno results were statically significant comparing to GloVe over the two ground truth datasets.

The results of the system were in general promising and can be applied not only to enrich and expand laymen vocabularies for medicine but any ontology for a domain, given an appropriate corpus for the domain. Our approach is applicable to narrow domains that may not have the huge training corpora typically used with word embedding approaches. In essence, by incorporating an external source of linguistic information, WordNet, and expanding the training corpus, we are getting more out of our training corpus. Our system can help building an application for patients where they can read their physician's letters more understandably and clearly. Moreover, the output of this system can be used to improve the results of healthcare search engines, entity recognition systems, and many others.

5.2 Future Work

For future work, we suggest further improving the GloVeSyno, GloVeHypo, GloVeHyper, GloVeIF algorithms. We also recommend using the output of GloVeSyno in different other applications such as search engine optimization or medical text translation. In our current research, we implemented our algorithms on only unigram seed terms. We plan to explore applying these algorithms to different word grams of different lengths. Also, we plan to treat the problem of word polysemy for a list of synonyms coming from the WordNet so that only synonyms that have very close meaning to the seed term will be added to the context of that term. Finally, we plan to implement our WordNet enriched corpus to state-of-the-art word embedding algorithms such as BERT [173], GPT-2 [174], and CTRL [175] algorithms.

6 References

[1] L. Répás, "Basics of Medical Terminology. Latin and Greek Origins," *Textbook for 1st Year Students of Medicine. Litográfia Nyomda, http://www. ilekt. med. unideb. hu/kiadvany/4latineng. pdf, accessed on,* vol. 20, no. 08, p. 2016, 2013.

[2] "Doctors told to use 'plain English,'" *BBC News*, Sep. 04, 2018.

[3] S. Vita *et al.*, "The 'Doctor Apollo' chatbot: a digital health tool to improve engagement of people living with HIV," in *JOURNAL OF THE INTERNATIONAL AIDS SOCIETY*, 2018, vol. 21.

[4] S. K. Mishra, D. Bharti, and N. Mishra, "Dr. Vdoc: A Medical Chatbot that Acts as a Virtual Doctor," *Research & Reviews: Journal of Medical Science and Technology*, vol. 6, no. 3, pp. 16–20, 2018.

[5] N. Rosruen and T. Samanchuen, "Chatbot Utilization for Medical Consultant System," in *2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, 2018, pp. 1–5.

[6] S. Fox, "Health Topics," *Pew Research Center: Internet, Science & Tech*, Feb. 01, 2011. https://www.pewinternet.org/2011/02/01/health-topics-3/ (accessed Oct. 21, 2019).

[7] G. T. Bosslet, A. M. Torke, S. E. Hickman, C. L. Terry, and P. R. Helft, "The patient–doctor relationship and online social networks: results of a national survey," *Journal of general internal medicine*, vol. 26, no. 10, pp. 1168–1174, 2011.

[8] D. R. George, L. S. Rovniak, and J. L. Kraschnewski, "Dangers and opportunities for social media in medicine," *Clinical obstetrics and gynecology*, vol. 56, no. 3, 2013.

[9] M. Modahl, L. Tompsett, and T. Moorhead, "Doctors, Patients & Social Media," *Social Media*, p. 16, 2011.

[10] C. Yeginsu, "Rx for British Doctors: Use Plain English Instead of Latin," *The New York Times*, Sep. 05, 2018.

[11] S. Blanchard, "Doctors told to write to their patients in plain English," *Daily Mail Online*, Sep. 04, 2018. http://www.dailymail.co.uk/health/article-6129323/Doctors-told-write-patients-plain-English.html (accessed Oct. 21, 2019).

[12] "Unified Medical Language System (UMLS)." https://www.nlm.nih.gov/research/umls/index.html (accessed Oct. 21, 2019).

[13] A. T. McCray and S. J. Nelson, "The representation of meaning in the UMLS," *Methods of information in medicine*, vol. 34, no. 01/02, pp. 193–201, 1995.

[14] S. EHOLIÉ, “Discovering Consumer Health Vocabulary from Patient-Generated Text in Social Media,” *SIFR Project - LIRMM*, Jun. 2016, [Online]. Available: https://wiki.lina.univ-nantes.fr/lib/exe/fetch.php?media=rapport_solene_eholie.pdf.

[15] Z. He, Z. Chen, S. Oh, J. Hou, and J. Bian, “Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach,” *Journal of biomedical informatics*, vol. 69, pp. 75–85, 2017.

[16] K. M. Doing-Harris and Q. Zeng-Treitler, “Computer-assisted update of a consumer health vocabulary through mining of social network data,” *Journal of medical Internet research*, vol. 13, no. 2, p. e37, 2011.

[17] N. Miller, E.-M. Lacroix, and J. E. B. Backus, “MEDLINEplus: building and maintaining the National Library of Medicine’s consumer health Web service,” *Bull Med Libr Assoc*, vol. 88, no. 1, pp. 11–17, Jan. 2000.

[18] M. Grüninger and M. S. Fox, “Methodology for the design and evaluation of ontologies,” 1995.

[19] M. Uschold and M. Gruninger, “Ontologies: Principles, methods and applications,” *The knowledge engineering review*, vol. 11, no. 2, pp. 93–136, 1996.

[20] R. Navigli, P. Velardi, and A. Gangemi, “Ontology learning and its application to automated terminology translation,” *IEEE Intelligent systems*, vol. 18, no. 1, pp. 22–31, 2003.

[21] S. Staab and R. Studer, *Handbook on ontologies*. Springer Science & Business Media, 2010.

[22] N. F. Noy and D. L. McGuinness, *Ontology development 101: A guide to creating your first ontology*. Stanford knowledge systems laboratory technical report KSL-01-05 and …, 2001.

[23] E. Zavitsanos, G. Paliouras, G. A. Vouros, and S. Petridis, “Learning subsumption hierarchies of ontology concepts from texts,” *Web Intelligence and Agent Systems: An International Journal*, vol. 8, no. 1, pp. 37–51, 2010.

[24] T. R. Gruber, “The role of common ontology in achieving sharable, reusable knowledge bases,” *KR*, vol. 91, pp. 601–602, 1991.

[25] R. Navigli, P. Velardi, and S. Faralli, “A graph-based algorithm for inducing lexical taxonomies from scratch,” 2011.

[26] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web,” *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[27] R. Navigli and S. P. Ponzetto, “BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network,” *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[28] L. M. Schriml *et al.*, “Disease Ontology: a backbone for disease semantic integration,” *Nucleic acids research*, vol. 40, no. D1, pp. D940–D946, 2011.

[29] M. Jarrar, *Building a Formal Arabic Ontology (Invited Paper). In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso, Arab League*. Tunis: sn, 2011.

[30] G. A. Miller, “WordNet: a lexical database for English,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[31] G. O. Consortium, “The gene ontology project in 2008,” *Nucleic acids research*, vol. 36, no. suppl_1, pp. D440–D444, 2007.

[32] O. Bodenreider, “The unified medical language system (UMLS): integrating biomedical terminology,” *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[33] “product description: personal health terminology (PHT).” Intelligent Medical Objects, Inc.

[34] C. E. Lipscomb, “Medical subject headings (MeSH),” *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[35] S. Gauch, J. Chaffee, and A. Pretschner, “Ontology-based personalized search and browsing,” p. 17.

[36] A. Pretschner and S. Gauch, “Ontology based personalized search,” in *Proceedings 11th International Conference on Tools with Artificial Intelligence*, Chicago, IL, USA, 1999, pp. 391–398, doi: 10.1109/TAI.1999.809829.

[37] J. Trajkova and S. Gauch, “Improving ontology-based user profiles,” in *Coupling approaches, coupling media and coupling languages for information retrieval*, 2004, pp. 380–390.

[38] J. Chaffee and S. Gauch, “Personal ontologies for web navigation,” in *Proceedings of the ninth international conference on Information and knowledge management*, 2000, pp. 227–234.

[39] V. Challam, S. Gauch, and A. Chandramouli, “Contextual search using ontology-based user profiles,” in *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, 2007, pp. 612–617.

[40] V. K. R. Challam and S. Gauch, “Contextual information retrieval using ontology based user profiles,” PhD Thesis, University of Kansas, Electrical Engineering and Computer Science, 2004.

[41] A. Maedche and S. Staab, “Ontology learning for the semantic web,” *IEEE Intelligent systems*, vol. 16, no. 2, pp. 72–79, 2001.

[42] D. Fensel, F. Van Harmelen, I. Horrocks, D. L. McGuinness, and P. F. Patel-Schneider, “OIL: An ontology infrastructure for the semantic web,” *IEEE intelligent systems*, vol. 16, no. 2, pp. 38–45, 2001.

[43] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, “Learning to map between ontologies on the semantic web,” in *Proceedings of the 11th international conference on World Wide Web*, 2002, pp. 662–673.

[44] S. A. McIlraith, T. C. Son, and H. Zeng, “Semantic web services,” *IEEE intelligent systems*, vol. 16, no. 2, pp. 46–53, 2001.

[45] V. Berrios *et al.*, “Cross-industry semantic interoperability, part three: The role of a top-level ontology,” Jul. 26, 2017. https://www.embedded-computing.com/embedded-computing-design/cross-industry-semantic-interoperability-part-three-the-role-of-a-top-level-ontology (accessed Sep. 18, 2019).

[46] M. R. Bautista-Zambrana, “Methodologies to build ontologies for terminological purposes,” *Procedia-Social and Behavioral Sciences*, vol. 173, pp. 264–269, 2015.

[47] T. R. Gruber, “Toward principles for the design of ontologies used for knowledge sharing?,” *International journal of human-computer studies*, vol. 43, no. 5–6, pp. 907–928, 1995.

[48] A. Hegazy, M. Sakre, and E. Khater, “Arabic Ontology model for financial Accounting,” *Procedia Computer Science*, vol. 62, pp. 513–520, 2015.

[49] “Apollo Home Page.” http://apollo.open.ac.uk/ (accessed Sep. 20, 2019).

[50] M. Weiten, “OntoSTUDIO® as a Ontology Engineering Environment,” in *Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies*, J. Davies, M. Grobelnik, and D. Mladenić, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 51–60.

[51] R. Alford, *ronwalf/swoop*. 2019.

[52] “protégé.” https://protege.stanford.edu/ (accessed Sep. 20, 2019).

[53] B. Fortuna, M. Grobelnik, and D. Mladenic, “OntoGen: Semi-automatic Ontology Editor,” in *Human Interface and the Management of Information. Interacting in Information Environments*, 2007, pp. 309–318.

[54] E. S. Alatrish, “Comparison some of ontology,” *Journal of Management Information Systems*, vol. 8, no. 2, pp. 018–024, 2013.

[55] E. D'Avanzo, A. Lieto, and T. Kuflik, "Manually vs semiautomatic domain specific ontology building," PhD Thesis, Thesis in Information and Commercial Electronics, 2008.

[56] D. FADL, S. ABBAS, and M. AREF, "AUTOMATIC ARABIC ONTOLOGY GENERATION FOR THE ANIMAL KINGDOM (NAAO).," *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 1, 2018.

[57] L. Zhou, "Ontology learning: state of the art and open issues," *Information Technology and Management*, vol. 8, no. 3, pp. 241–252, 2007.

[58] M. Hazman, S. R. El-Beltagy, and A. Rafea, "A survey of ontology learning approaches," *International Journal of Computer Applications*, vol. 22, no. 9, pp. 36–43, 2011.

[59] A. Gómez-Pérez and D. Manzano-Macho, "A survey of ontology learning methods and techniques," *OntoWeb Deliverable D*, vol. 1, no. 5, 2003.

[60] J. Ronk, "Structured, semi structured and unstructured data," *Jeremy Ronk*, Sep. 01, 2014. https://jeremyronk.wordpress.com/2014/09/01/structured-semi-structured-and-unstructured-data/ (accessed Sep. 19, 2019).

[61] C. Faria, I. Serra, and R. Girardi, "A domain-independent process for automatic ontology population from text," *Science of Computer Programming*, vol. 95, pp. 26–43, Dec. 2014, doi: 10.1016/j.scico.2013.12.005.

[62] I. Antonov, I. Bruttan, D. Andreev, and L. Motaylenko, "The Method of Automated Building of Domain Ontology," in *Proceedings of the 12th International Scientific and Practical Conference. Volume II*, 2019, vol. 34, p. 37.

[63] M. Li, X.-Y. Du, and S. Wang, "Learning ontology from relational database," in *2005 International Conference on Machine Learning and Cybernetics*, 2005, vol. 6, pp. 3410–3415.

[64] L. Zhang and J. Li, "Automatic generation of ontology based on database," *Journal of Computational Information Systems*, vol. 7, no. 4, pp. 1148–1154, 2011.

[65] Z. Lin, R. Lu, Y. Xiong, and Y. Zhu, "Learning ontology automatically using topic model," in *2012 International Conference on Biomedical Engineering and Biotechnology*, 2012, pp. 360–363.

[66] D. Alfonso-Hermelo, P. Langlais, and L. Bourg, "Automatically Learning a Human-Resource Ontology from Professional Social-Network Data," in *Canadian Conference on Artificial Intelligence*, 2019, pp. 132–145.

[67] A. Arora, "Automatic Ontology Construction: Ontology From Plain Text Using Conceptualization and Semantic Roles," in *Critical Approaches to Information Retrieval Research*, IGI Global, 2020, pp. 109–149.

[68] A. Alobaid, D. Garijo, M. Poveda-Villalón, I. Santana-Perez, A. Fernández-Izquierdo, and O. Corcho, "Automating ontology engineering support activities with OnToology," *Journal of Web Semantics*, vol. 57, p. 100472, 2019.

[69] D. B. Hier and S. U. Brint, "A Neuro-ontology for the neurological examination," *BMC Med Inform Decis Mak*, vol. 20, no. 1, p. 47, Mar. 2020, doi: 10.1186/s12911-020-1066-7.

[70] H. Yilahun, S. Imam, and A. Hamdulla, "Ontology expansion based on UWN reusability," *International Journal of Information and Communication Technology*, vol. 16, no. 4, pp. 339–352, Jan. 2020, doi: 10.1504/IJICT.2020.107588.

[71] "Yahoo," *Yahoo*. http://www.yahoo.com (accessed Sep. 24, 2019).

[72] E. Agirre, O. Ansa, E. Hovy, and D. Martinez, "Enriching very large ontologies using the WWW," *arXiv:cs/0010026*, Oct. 2000, Accessed: Sep. 24, 2019. [Online]. Available: http://arxiv.org/abs/cs/0010026.

[73] H. P. Luong, S. Gauch, and Q. Wang, "Ontology learning through focused crawling and information extraction," in *2009 International Conference on Knowledge and Systems Engineering*, 2009, pp. 106–112.

[74] H. Luong, S. Gauch, and Q. Wang, "Ontology learning using word net lexical expansion and text mining," *Theory and Applications for Advanced Text Mining*, p. 101, 2012.

[75] Q. Wang, S. Gauch, and H. Luong, "Ontology concept enrichment via text mining," in *IADIS international conference on internet technologies & society*, 2010, pp. 147–154.

[76] H. P. Luong, S. Gauch, and M. Speretta, "Enriching concept descriptions in an amphibian ontology with vocabulary extracted from wordnet," in *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*, 2009, pp. 1–6.

[77] H. P. Luong, S. Gauch, and Q. Wang, "Ontology-based focused crawling," in *2009 International Conference on Information, Process, and Knowledge Management*, 2009, pp. 123–128.

[78] H. Luong, S. Gauch, Q. Wang, and A. Maglia, "An ontology learning framework using focused crawler and text mining," *International Journal On Advances in Life Sciences*, vol. 1, no. 2, pp. 99–109, 2009.

[79] M. Speretta and S. Gauch, "Using text mining to enrich the vocabulary of domain ontologies," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, 2008, pp. 549–552.

[80] M. Ali, S. Fathalla, S. Ibrahim, M. Kholief, and Y. F. Hassan, "CLOE: a cross-lingual ontology enrichment using multi-agent architecture," *Enterprise Information Systems*, vol. 13, no. 7–8, pp. 1002–1022, Sep. 2019, doi: 10.1080/17517575.2019.1592232.

[81] N. Ye, A. Pudhiyaveetil, and S. Gauch, “Mining Hidden Concepts for Ontology Extension Using Multivariate Probabilistic Modeling,” in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2009, vol. 1, pp. 371–374.

[82] “The 2012 ACM Computing Classification System.” https://www.acm.org/publications/class-2012 (accessed Oct. 15, 2019).

[83] “CiteSeerX.” https://citeseerx.ist.psu.edu/index;jsessionid=611FF4CDE32CA191B521F964DD69044A (accessed Oct. 15, 2019).

[84] S. Peroni and D. Shotton, “The SPAR ontologies,” in *International Semantic Web Conference*, 2018, pp. 119–136.

[85] M. del C. S. de Figueroa, B. Advisors, A. G. Pérez, and M. F. López, “NeOn methodology for building ontology networks: Specification, scheduling and reuse,” PhD Thesis, PhD thesis, Universidad Politécnica de Madrid. Facultad de Informática …, 2010.

[86] M. Poveda-Villalón, A. Gómez-Pérez, and M. C. Suárez-Figueroa, “Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation,” *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 10, no. 2, pp. 7–34, 2014.

[87] M. Tapia-Leon, I. Santana-Perez, M. Poveda-Villalón, P. Espinoza-Arias, J. Chicaiza, and O. Corcho, “Extension of the BiDO Ontology to Represent Scientific Production,” in *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, 2019, pp. 166–172.

[88] A. L. Rector, R. Qamar, and T. Marley, “Binding ontologies and coding systems to electronic health records and messages,” *Applied Ontology*, vol. 4, no. 1, pp. 51–69, 2009.

[89] K. Donnelly, “SNOMED-CT: The advanced terminology and coding system for eHealth,” *Studies in health technology and informatics*, vol. 121, p. 279, 2006.

[90] “What is a medical terminology, taxonomy, or ontology?” https://www.lexigram.io/lexipedia/medical-taxonomy-terminology-hierarchy/ (accessed Oct. 17, 2019).

[91] H. J. Lowe and G. O. Barnett, “MicroMeSH: A Microcomputer System for Searching and Exploring the National Library of Medicine’s Medical Subject Headings (MeSH) Vocabulary,” *Proc Annu Symp Comput Appl Med Care*, pp. 717–720, Nov. 1987.

[92] T. Nakazato, H. Bono, H. Matsuda, and T. Takagi, “Gendoo: Functional profiling of gene and disease features using MeSH vocabulary,” *Nucleic Acids Res*, vol. 37, no. suppl_2, pp. W166–W169, Jul. 2009, doi: 10.1093/nar/gkp483.

[93] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, "DISEASES: Text mining and data integration of disease–gene associations," *Methods*, vol. 74, pp. 83–89, Mar. 2015, doi: 10.1016/j.ymeth.2014.11.020.

[94] M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña-López, "Query expansion with a medical ontology to improve a multimodal information retrieval system," *Computers in Biology and Medicine*, vol. 39, no. 4, pp. 396–403, Apr. 2009, doi: 10.1016/j.compbiomed.2009.01.012.

[95] M. Stevenson and Y. Guo, "Disambiguation of ambiguous biomedical terms using examples generated from the UMLS Metathesaurus," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 762–773, Oct. 2010, doi: 10.1016/j.jbi.2010.06.001.

[96] G. Dennis *et al.*, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biology*, vol. 4, no. 9, p. R60, Aug. 2003, doi: 10.1186/gb-2003-4-9-r60.

[97] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat Protoc*, vol. 4, no. 1, pp. 44–57, Jan. 2009, doi: 10.1038/nprot.2008.211.

[98] H. Attrill *et al.*, "Annotation of gene product function from high-throughput studies using the Gene Ontology," *Database (Oxford)*, vol. 2019, Jan. 2019, doi: 10.1093/database/baz007.

[99] H. Park and N. Hardiker, "Clinical terminologies: a solution for semantic interoperability," *Journal of Korean Society of Medical Informatics*, vol. 15, no. 1, pp. 1–11, 2009.

[100] Q. Zeng, S. Kogan, N. Ash, and R. A. Greenes, "Patient and clinician vocabulary: how different are they?," *Medinfo*, vol. 10, no. Pt 1, pp. 399–403, 2001.

[101] "Italian Consumer-oriented Medical Vocabulary (ICMV) | E-HEALTH." https://ehealth.fbk.eu/resources/italian-consumer-oriented-medical-vocabulary-icmv (accessed Sep. 17, 2019).

[102] P. D. Marshall, "Bridging the terminology gap between health care professionals and patients with the Consumer Health Terminology (CHT)," in *Proceedings of the AMIA Symposium*, 2000, p. 1082.

[103] L. Hou, H. Kang, Y. Liu, L. Li, and J. Li, "Mining and standardizing chinese consumer health terms," *BMC medical informatics and decision making*, vol. 18, no. 5, p. 120, 2018.

[104] M. Adnan, J. Warren, and M. Orr, "SemLink — Dynamic generation of hyperlinks to enhance patient readability of discharge summaries," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, Porto, Portugal, Jun. 2013, pp. 35–40, doi: 10.1109/CBMS.2013.6627761.

[105] S. Kandula, D. Curtis, and Q. Zeng-Treitler, "A semantic and syntactic text simplification tool for health content," in *AMIA annual symposium proceedings*, 2010, vol. 2010, p. 366.

[106] Q. Zeng-Treitler, S. Goryachev, H. Kim, A. Keselman, and D. Rosendale, "Making texts in electronic health records comprehensible to consumers: a prototype translator," in *AMIA Annual Symposium Proceedings*, 2007, vol. 2007, p. 846.

[107] B. Qenam, T. Y. Kim, M. J. Carroll, and M. Hogarth, "Text Simplification Using Consumer Health Vocabulary to Generate Patient-Centered Radiology Reporting: Translation and Evaluation," *Journal of Medical Internet Research*, vol. 19, no. 12, p. e417, 2017, doi: 10.2196/jmir.8536.

[108] R. D. Zielstorff, "Controlled vocabularies for consumer health," *Journal of biomedical informatics*, vol. 36, no. 4–5, pp. 326–333, 2003.

[109] K. O'Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, and G. Gonzalez, "Pharmacovigilance on twitter? Mining tweets for adverse drug reactions," in *AMIA annual symposium proceedings*, 2014, vol. 2014, p. 924.

[110] H. Yang and C. C. Yang, "Mining a Weighted Heterogeneous Network Extracted from Healthcare-Specific Social Media for Identifying Interactions between Drugs," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, Atlantic City, NJ, USA, Nov. 2015, pp. 196–203, doi: 10.1109/ICDMW.2015.73.

[111] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 671–681, 2015.

[112] H. Yang and C. C. Yang, "Using Health-Consumer-Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 4, p. 55:1-55:27, Jul. 2015, doi: 10.1145/2700482.

[113] "DiseaseOntology/HumanDiseaseOntology," *GitHub*. https://github.com/DiseaseOntology/HumanDiseaseOntology (accessed Sep. 14, 2019).

[114] M. Ashburner *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[115] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, and G. A. Landrum, "Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter," *Journal of medicinal chemistry*, vol. 59, no. 9, pp. 4385–4402, 2016.

[116] V. Law *et al.*, "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic acids research*, vol. 42, no. D1, pp. D1091–D1097, 2013.

[117] S. Köhler *et al.*, "The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data," *Nucleic acids research*, vol. 42, no. D1, pp. D966–D974, 2013.

[118] "NCBO BioPortal." https://bioportal.bioontology.org/ontologies (accessed Oct. 18, 2019).

[119] J. Huang *et al.*, "OMIT: Dynamic, Semi-Automated Ontology Development for the microRNA Domain," *PLOS ONE*, vol. 9, no. 7, p. e100855, Jul. 2014, doi: 10.1371/journal.pone.0100855.

[120] "MedlinePlus - Health Information from the National Library of Medicine." https://medlineplus.gov/ (accessed Sep. 23, 2019).

[121] Q. T. Zeng, T. Tse, J. Crowell, G. Divita, L. Roth, and A. C. Browne, "Identifying Consumer-Friendly Display (CFD) Names for Health Concepts," *AMIA Annu Symp Proc*, vol. 2005, pp. 859–863, 2005.

[122] Q. T. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 24–29, 2006.

[123] Y. He *et al.*, "CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis," *Scientific Data*, vol. 7, no. 1, Art. no. 1, Jun. 2020, doi: 10.1038/s41597-020-0523-6.

[124] K. Doing-Harris, Y. Livnat, and S. Meystre, "Automated concept and relationship extraction for the semi-automated ontology management (SEAM) system," *J Biomed Semantics*, vol. 6, Apr. 2015, doi: 10.1186/s13326-015-0011-7.

[125] M. B. do Amaral, A. Roberts, and A. L. Rector, "NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs.," in *Proceedings of the AMIA Symposium*, 2000, p. 76.

[126] V. Kashyap, C. Ramakrishnan, and T. Rindflesch, "Towards (Semi-) Automatic Generation of Bio-Medical Ontologies," *AMIA Annual Symposium Proceedings, 2003*, p. 886, Nov. 2003.

[127] D. Sánchez and A. Moreno, *Learning medical ontologies from the Web*. .

[128] W. Zhou *et al.*, "A Semi-automatic Ontology Learning Based on WordNet and Event-based Natural Language Processing," in *2006 International Conference on Information and Automation*, Dec. 2006, pp. 240–244, doi: 10.1109/ICINFA.2006.374119.

[129] C. Pesquita and F. M. Couto, "Predicting the Extension of Biomedical Ontologies," *PLOS Computational Biology*, vol. 8, no. 9, p. e1002630, Sep. 2012, doi: 10.1371/journal.pcbi.1002630.

[130] K. Frantzi, S. Ananiadou, and H. Mima, “Automatic recognition of multi-word terms:. the c-value/nc-value method,” *International journal on digital libraries*, vol. 3, no. 2, pp. 115–130, 2000.

[131] Q. T. Zeng *et al.*, “Term Identification Methods for Consumer Health Vocabulary Development,” *J Med Internet Res*, vol. 9, no. 1, Mar. 2007, doi: 10.2196/jmir.9.1.e4.

[132] Q. Zheng and X.-J. Wang, “GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis,” *Nucleic Acids Res*, vol. 36, no. suppl_2, pp. W358–W363, Jul. 2008, doi: 10.1093/nar/gkn276.

[133] N. Shanavas, H. Wang, Z. Lin, and G. Hawe, “Ontology-based enriched concept graphs for medical document classification,” *Information Sciences*, vol. 525, pp. 172–181, Jul. 2020, doi: 10.1016/j.ins.2020.03.006.

[134] J. Baldridge, “The openNLP project.” Jan. 18, 2005, [Online]. Available: http://opennlp.apache.org/index.

[135] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[136] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[137] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[138] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[139] “WHO | International Classification of Diseases (ICD) Information Sheet,” *WHO*. https://www.who.int/classifications/icd/factsheet/en/ (accessed Oct. 02, 2019).

[140] G. Gu *et al.*, “Development of a Consumer Health Vocabulary by Mining Health Forum Texts Based on Word Embedding: Semiautomatic Approach,” *JMIR Medical Informatics*, vol. 7, no. 2, p. e12704, 2019, doi: 10.2196/12704.

[141] A. Rai, R, and M. Learning, “What is Text Mining: Techniques and Applications,” *upGrad blog*, Jun. 01, 2019. https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/ (accessed Oct. 19, 2019).

[142] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[143] M. W. Berry and M. Castellanos, “Survey of text mining,” *Computing Reviews*, vol. 45, no. 9, p. 548, 2004.

[144] M. M. Deza and E. Deza, “Encyclopedia of distances,” in *Encyclopedia of distances*, Springer, 2009, pp. 1–583.

[145] T. Mikolov, W. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 746–751.

[146] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[147] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[148] S. Gauch, J. Wang, and S. M. Rachakonda, “A corpus analysis approach for automatic query expansion and its extension to multiple databases,” *ACM Transactions on Information Systems (TOIS)*, vol. 17, no. 3, pp. 250–269, 1999.

[149] S. A. Hasan and O. Farri, “Clinical Natural Language Processing with Deep Learning,” in *Data Science for Healthcare*, S. Consoli, D. Reforgiato Recupero, and M. Petković, Eds. Cham: Springer International Publishing, 2019, pp. 147–171.

[150] J. A. Minarro-Giménez, O. Marin-Alonso, and M. Samwald, “Exploring the application of deep learning techniques on medical text corpora.,” *Studies in health technology and informatics*, vol. 205, pp. 584–588, 2014.

[151] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, “Medical text classification using convolutional neural networks,” *Stud Health Technol Inform*, vol. 235, pp. 246–250, 2017.

[152] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza, “Medical semantic similarity with a neural language model,” in *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 2014, pp. 1819–1822.

[153] J. A. Minarro-Giménez, O. Marín-Alonso, and M. Samwald, “Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation,” *arXiv preprint arXiv:1502.03682*, 2015.

[154] C. Wang, L. Cao, and B. Zhou, “Medical synonym extraction with concept space models,” 2015.

[155] T. Kenter and M. De Rijke, “Short text similarity with word embeddings,” in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1411–1420.

[156] R. Brochier, A. Guille, and J. Velcin, "Global vectors for node representations," in *The World Wide Web Conference*, 2019, pp. 2587–2593.

[157] A. George, B. G. HB, and K. P. Soman, "Teamcen at semeval-2018 task 1: global vectors representation in emotion detection," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 334–338.

[158] M. Espinoza, A. Gómez-Pérez, and E. Mena, "Enriching an ontology with multilingual information," in *European Semantic Web Conference*, 2008, pp. 333–347.

[159] R. Navigli and P. Velardi, "Enriching a formal ontology with a thesaurus: an application in the cultural heritage domain," in *Proceedings of the 2nd workshop on ontology learning and population: Bridging the gap between text and knowledge*, 2006, pp. 1–9.

[160] M. Warin, H. Oxhammar, and M. Volk, "Enriching an ontology with wordnet based on similarity measures," 2005.

[161] H. Kilicoglu *et al.*, "Semantic annotation of consumer health questions," *BMC bioinformatics*, vol. 19, no. 1, p. 34, 2018.

[162] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, and V. Malykh, "Medical concept normalization in social media posts with recurrent neural networks," *Journal of biomedical informatics*, vol. 84, pp. 93–102, 2018.

[163] "MedHelp - Medical Information, Forums and Communities." https://www.medhelp.org/about (accessed Jul. 09, 2019).

[164] M. F. Porter, "An algorithm for suffix stripping," *Program*, 2006.

[165] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

[166] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[167] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint cmp-lg/9511007*, 1995.

[168] M. S. Park, Z. He, Z. Chen, S. Oh, and J. Bian, "Consumers' use of UMLS concepts on social media: diabetes-related textual data analysis in blog and social Q&A sites," *JMIR medical informatics*, vol. 4, no. 4, p. e41, 2016.

[169] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.

[170] E. M. Voorhees and D. K. Harman, "The Eighth Text REtrieval Conference (TREC-8)," Nov. 2000, Accessed: Oct. 29, 2020. [Online]. Available: https://www.nist.gov/publications/eighth-text-retrieval-conference-trec-8.

[171] "GloVe: Global Vectors for Word Representation." https://nlp.stanford.edu/projects/glove/ (accessed Dec. 17, 2020).

[172] "Common Crawl." https://commoncrawl.org/ (accessed Dec. 17, 2020).

[173] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019, Accessed: Mar. 18, 2020. [Online]. Available: http://arxiv.org/abs/1810.04805.

[174] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," p. 24, 2019.

[175] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A Conditional Transformer Language Model for Controllable Generation," *arXiv:1909.05858 [cs]*, Sep. 2019, Accessed: Feb. 08, 2021. [Online]. Available: http://arxiv.org/abs/1909.05858.